

Computergestützte Inhaltsanalyse von digitalen Videoarchiven

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
der Universität Mannheim

vorgelegt von

Dipl.-Wirtsch.-Inf. Stephan Kopf
aus Mannheim

Mannheim, 2006

Dekan:	Professor Dr. M. Krause, Universität Mannheim
Referent:	Professor Dr. W. Effelsberg, Universität Mannheim
Korreferent:	Professor Dr. R. Lienhart, Universität Augsburg

Tag der mündlichen Prüfung: 1. März 2007

Zusammenfassung

Der Übergang von analogen zu digitalen Videos hat in den letzten Jahren zu großen Veränderungen innerhalb der Filmarchive geführt. Insbesondere durch die Digitalisierung der Filme ergeben sich neue Möglichkeiten für die Archive. Eine Abnutzung oder Alterung der Filmrollen ist ausgeschlossen, so dass die Qualität unverändert erhalten bleibt. Zudem wird ein netzbasierter und somit deutlich einfacherer Zugriff auf die Videos in den Archiven möglich. Zusätzliche Dienste stehen den Archivaren und Anwendern zur Verfügung, die erweiterte Suchmöglichkeiten bereitstellen und die Navigation bei der Wiedergabe erleichtern. Die Suche innerhalb der Videoarchive erfolgt mit Hilfe von Metadaten, die weitere Informationen über die Videos zur Verfügung stellen. Ein großer Teil der Metadaten wird manuell von Archivaren eingegeben, was mit einem großen Zeitaufwand und hohen Kosten verbunden ist.

Durch die computergestützte Analyse eines digitalen Videos ist es möglich, den Aufwand bei der Erzeugung von Metadaten für Videoarchive zu reduzieren. Im ersten Teil dieser Dissertation werden neue Verfahren vorgestellt, um wichtige semantische Inhalte der Videos zu erkennen. Insbesondere werden neu entwickelte Algorithmen zur Erkennung von Schnitten, der Analyse der Kamerabewegung, der Segmentierung und Klassifikation von Objekten, der Texterkennung und der Gesichtserkennung vorgestellt.

Die automatisch ermittelten semantischen Informationen sind sehr wertvoll, da sie die Arbeit mit digitalen Videoarchiven erleichtern. Die Informationen unterstützen nicht nur die Suche in den Archiven, sondern führen auch zur Entwicklung neuer Anwendungen, die im zweiten Teil der Dissertation vorgestellt werden. Beispielsweise können computergenerierte Zusammenfassungen von Videos erzeugt oder Videos automatisch an die Eigenschaften eines Abspielgerätes angepasst werden.

Ein weiterer Schwerpunkt dieser Dissertation liegt in der Analyse historischer Filme. Vier europäische Filmarchive haben eine große Anzahl historischer Videodokumentationen zur Verfügung gestellt, welche Anfang bis Mitte des letzten Jahrhunderts gedreht und in den letzten

Jahren digitalisiert wurden. Durch die Lagerung und Abnutzung der Filmrollen über mehrere Jahrzehnte sind viele Videos stark verrauscht und enthalten deutlich sichtbare Bildfehler. Die Bildqualität der historischen Schwarz-Weiß-Filme unterscheidet sich signifikant von der Qualität aktueller Videos, so dass eine verlässliche Analyse mit bestehenden Verfahren häufig nicht möglich ist. Im Rahmen dieser Dissertation werden neue Algorithmen vorgestellt, um eine zuverlässige Erkennung von semantischen Inhalten auch in historischen Videos zu ermöglichen.

Abstract

The change from analog to digital videos in recent years has led to significant improvements in film archives. New possibilities for the archives arise due to the digitalization of films and videos. Wear-out and aging of film reels can be eliminated and a long-term preservation of the quality will be guaranteed. Additionally, the net-based access is much easier and faster than the manual transport of film reels. New services for archivists and users are available which enable new search possibilities and facilitate fast and efficient navigation during the playback of videos. Metadata provide additional information about the content of videos and support the search within the archives. In spite of the time exposure and high costs, a large part of the metadata is manually added by the archivists.

The automatic analysis of digital video archives reduces the effort to create metadata significantly. Presented in the first part of this dissertation are new algorithms and techniques to identify and extract relevant semantic content in videos. In particular, new algorithms were developed to detect shot boundaries in videos, to analyze the camera motion, to segment and classify moving objects in videos, to perform optical character recognition, and to detect and recognize faces in videos.

Automatically extracted semantic information is very valuable due to the fact that this information supports the work with digital archives. The additional information not only enables the search of videos within an archive but also leads to new applications, which are presented in the second part of this dissertation. Two sample applications are examined: automatically generated video summaries and video adaptation algorithms which enable the playback of videos on arbitrary devices.

Another focal point of this dissertation is the analysis of historical films. Four European film archives provided a vast number of historical video documentaries stemming from the beginning to the middle of the last century. The storage and wear-out of the film reels over several decades led to noisy videos and a large number of errors in the images. The quality of the hi-

historical black-and-white films is significantly lower than that of current videos, and a reliable analysis with existing techniques is often not possible. New algorithms are presented in this dissertation which enable the identification of semantic content even in historical videos.

Vorwort

Die vorliegende Arbeit entstand während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Lehrstuhl für Praktische Informatik IV der Universität Mannheim.

Ganz besonders herzlich möchte ich Herrn Prof. Dr. Wolfgang Effelsberg für die Unterstützung bei der Entstehung der Arbeit, den Hinweisen und Denkanstößen, sowie der konstruktiven Kritik danken. Auch für die Möglichkeit, mich jederzeit mit Fragen an ihn wenden und viele internationale wissenschaftliche Konferenzen besuchen zu können, möchte ich mich herzlich bedanken.

Herrn Prof. Dr. Rainer Lienhart danke ich für die Übernahme des Korreferats.

Wesentliche Ideen und Algorithmen zur Objekterkennung und der automatischen Erzeugung von Zusammenfassungen für Videos sind im Rahmen des Projektes *European Chronicles Online* entstanden. Insbesondere den Archiven Instituto Luce (Italien), Memoriav (Schweiz), Netherlands Institute for Sound and Vision (Niederlande) und Institut Nationale de l’Audiovisuel (Frankreich), die umfangreiche Sammlungen mit historischen Videodokumentationen zur Verfügung gestellt haben und mit denen eine enge Zusammenarbeit erfolgte, möchte ich danken. Ein weiterer Schwerpunkt meiner wissenschaftlichen Tätigkeit war die Verbesserung der Lehre durch den Einsatz mobiler Geräte und die Positionsbestimmung innerhalb von Gebäuden. Obwohl zahlreiche Publikationen in diesen Bereichen entstanden sind [6, 263, 264, 265, 266, 267, 268, 282, 283, 288, 289, 296, 547], werden diese innerhalb der Arbeit wegen ihrer deutlichen thematischen Abweichung nicht weiter berücksichtigt. Im Rahmen der mit meiner wissenschaftlichen Tätigkeit verbundenen Projekte möchte ich dem Learning Lab Lower Saxony (L3S), dem Wallenberg Global Learning Network (WGLN), der Landesstiftung Baden-Württemberg, dem Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg und der Deutschen Forschungsgemeinschaft danken.

Meinen aktuellen und ehemaligen Kollegen Marcel Busse, Holger Füßler, Thomas Haenselmann, Thomas King, Christoph Kuhmünch, Christian Liebig, Fleming Lampi, Martin Mauve,

Nicolai Scheele, Claudia Schremmer, Matthias Transier und Jürgen Vogel möchte ich für die gute und freundschaftliche Zusammenarbeit und die Möglichkeit danken, Ideen gemeinsam zu diskutieren. Auch danke ich zahlreichen Diplomanden, Studienarbeitern und wissenschaftlichen Hilfskräften.

Ganz besonderer Dank gilt Gerald Kühne, der mich zu Beginn meiner Arbeit wesentlich unterstützt hat, sowie Dirk Farin, der immer bereit war, Ideen gemeinsam zu diskutieren. Weiterer Dank gilt unserem Systemadministrator Walter Müller, unserer Sekretärin Ursula Eckle und unserer ehemaligen Sekretärin Betty Weyerer.

Der größte Dank gilt meiner Frau Stephanie, die meine Arbeit korrekturgelesen und mir Freiräume für meine Forschung geschaffen hat, indem sie ihre Arbeitsstelle reduzierte und sich um unsere Tochter Amelie kümmerte. Auch meiner Schwiegermutter, die in unserer Abwesenheit Amelie ganz lieb umsorgt, und meinen Eltern, die uns häufig unterstützt haben, gilt herzlicher Dank.

Inhalt

Abbildungsverzeichnis	XIII
Tabellenverzeichnis	XVII
1 Einleitung	1
I Algorithmen zur automatischen Analyse von Videos	5
2 Erkennung von Schnitten in Videos	7
2.1 Klassifikation eines Schnittes	8
2.2 Computergestützte Erkennung eines Schnittes	11
2.2.1 Pixelbasierte Verfahren zur Schnitterkennung	13
2.2.2 Schnitterkennung mit Histogrammen	14
2.2.3 Schnitterkennung durch Analyse der Standardabweichung	17
2.2.4 Kantenbasierte Verfahren zur Schnitterkennung	19
2.2.5 Verbesserung der Schnitterkennung durch Bewegungsanalyse	22
2.3 Experimentelle Ergebnisse	22
2.3.1 Theoretische Obergrenzen für die Erkennung harter Schnitte	24
2.3.2 Optimierungen zur Erkennung harter Schnitte	25
2.3.3 Theoretische Obergrenzen für die Erkennung weicher Schnitte	28
2.3.4 Optimierungen zur Erkennung weicher Schnitte	29
2.3.5 Klassifikationsergebnisse für harte und weiche Schnitte	31
2.3.6 Schnitterkennung in historischen Videos	33
2.4 Zusammenfassung	37

3	Analyse der Kamerabewegung	39
3.1	Modellierung der Kamerabewegung	40
3.2	Berechnung von Bewegungsvektoren	41
3.3	Schätzung der Parameter des Kameramodells	42
3.4	Exakte Berechnung des Kameramodells	45
3.5	Experimentelle Ergebnisse	47
3.6	Zusammenfassung	52
4	Objektsegmentierung durch Bewegungsanalyse	55
4.1	Kamerabewegungen zwischen beliebigen Bildern	56
4.2	Transformation eines Bildes	57
4.3	Konstruktion von Hintergrundbildern	59
4.4	Segmentierung von Objekten	63
4.5	Experimentelle Ergebnisse	66
4.6	Zusammenfassung	70
5	Klassifikation von Objekten	71
5.1	Parametrisierung der Kontur	74
5.2	Globale geometrische Konturdeskriptoren	75
5.3	Krümmungsbasierter Skalenraum	76
5.4	Abbildungen im krümmungsbasierten Skalenraum	77
5.5	Vergleich von Konturen	79
5.5.1	Rotationsinvarianter Konturvergleich	81
5.5.2	Merkmale der Skalenraumabbildungen	82
5.6	Vermeidung von Mehrdeutigkeiten	83
5.7	Klassifikation konvexer Objektregionen	84
5.8	Aggregation der Klassifikationsergebnisse für Videosequenzen	88
5.8.1	Anzahl erkannter Objektklassen	88
5.8.2	Aggregation über die Distanz zur Objektklasse	89
5.9	Experimentelle Ergebnisse	90
5.9.1	Objekte der Datenbank	90
5.9.2	Testsequenzen zur Objekterkennung	92
5.9.3	Klassifikation mit Hilfe der Merkmale des krümmungsbasierten Skalenraums	92
5.9.4	Erweiterung des Skalenraumvergleichs durch zusätzliche Merkmale	95

5.9.5	Klassifikation mit transformierten Konturen	99
5.9.6	Objekterkennung in historischen Videos	100
5.10	Zusammenfassung	103
6	Erkennung von Textregionen und Buchstaben	105
6.1	Existierende Verfahren zur Texterkennung	107
6.2	Erkennung von Textregionen	109
6.3	Segmentierung von Buchstaben	110
6.3.1	Ermittlung der Trenner zwischen Buchstaben	110
6.3.2	Identifikation der Textpixel	113
6.4	Klassifikation von Buchstaben	116
6.5	Analyse der Klassifikationsergebnisse	118
6.5.1	Erkennung von Buchstaben ohne Segmentierungsfehler	118
6.5.2	Vergleich bei fehlerhafter Segmentierung	119
6.5.3	Texterkennung in Bildern und Videos	120
6.6	Zusammenfassung	124
7	Gesichtserkennung	125
7.1	Anforderungen an Algorithmen zur Gesichtserkennung	126
7.2	Verfahren zur Gesichtserkennung	127
7.2.1	Modellbasierte Verfahren	128
7.2.2	Konnektionistische Verfahren	130
7.3	Lokalisierung und Erkennung von Gesichtern in Videos	134
7.3.1	Lokalisierung von Gesichtsregionen	134
7.3.2	Segmentierung eines Gesichtes	135
7.3.3	Klassifikation eines Gesichtes	138
7.4	Experimentelle Ergebnisse	139
7.5	Zusammenfassung	142
II	Anwendungen zur Analyse digitaler Videoarchive	143
8	Adaption von Videos	145
8.1	Verfahren zur Adaption multimedialer Inhalte	147
8.1.1	Unterstützung der Adaption durch Standardisierungsverfahren	148

8.1.2	Verfahren zur Adaption von Videos	149
8.2	Anpassung der Farbtiefe eines Videos	151
8.3	Anpassung der Bildauflösung eines Videos	156
8.3.1	Identifikation der semantischen Merkmale in Videos	157
8.3.2	Bewertung eines semantischen Merkmals	158
8.3.3	Auswahl und Kombination von Bildregionen	160
8.3.4	Festlegung der Regionen für Kameraeinstellungen	162
8.4	Anpassung der Bildqualität historischer Videos	164
8.4.1	Korrektur der Helligkeit in historischen Videos	164
8.4.2	Korrektur von Streifen und Kratzern im Bild	165
8.4.3	Korrektur verwackelter Kameraeinstellungen	166
8.5	Experimentelle Ergebnisse	168
8.6	Zusammenfassung	172
9	Computergenerierte Zusammenfassungen von Videos	173
9.1	Heuristiken zur Erzeugung von Zusammenfassungen	175
9.1.1	Allgemeine Merkmale zur Beschreibung von Kameraeinstellungen	176
9.1.2	Genrespezifische Merkmale zur Auswahl von Kameraeinstellungen	178
9.1.3	Statische Zusammenfassungen von Videos	179
9.1.4	Dynamische Zusammenfassungen von Videos	180
9.2	Systemüberblick	181
9.3	Strukturelle und semantische Analyse des Videos	182
9.3.1	Schnitterkennung und Auswahl repräsentativer Einzelbilder	183
9.3.2	Gruppierung ähnlicher Kameraeinstellungen	184
9.3.3	Erkennung von Szenen	185
9.3.4	Kamerabewegung	186
9.3.5	Bewegungsaktivität	187
9.3.6	Gesichter und Objekte	187
9.3.7	Analyse des Audiosignals	188
9.4	Auswahl relevanter Kameraeinstellungen	188
9.4.1	Bewertung der Kamerabewegung	189
9.4.2	Bewertung der Bewegungsaktivität	191
9.4.3	Bewertung der Gesichter und Objekte	191
9.4.4	Bewertung des Kontrastes	192

9.4.5	Bewertung der Ähnlichkeit von Kameraeinstellungen	192
9.4.6	Bewertung der Szenen	192
9.4.7	Bewertung der Verteilung der Kameraeinstellungen	193
9.5	Erzeugung einer Zusammenfassung	194
9.5.1	Auswahl von Kameraeinstellungen	194
9.5.2	Überprüfung der ausgewählten Kameraeinstellungen	195
9.5.3	Speicherung der Zusammenfassung	197
9.6	Experimentelle Ergebnisse	197
9.6.1	Statische Zusammenfassungen von Videos	198
9.6.2	Dynamische Zusammenfassungen von Videos	201
9.7	Zusammenfassung	202
10	Analyse der Bewegungen von Objekten und Personen	205
10.1	Verfahren zur Analyse von Bewegungen	207
10.2	Systemüberblick	208
10.3	Erweiterung der Datenbank	209
10.4	Aggregation der Klassifikationsergebnisse	210
10.5	Semantische Analyse der Fahrt eines PKWs	212
10.6	Semantische Analyse der Bewegung einer Person	215
10.7	Zusammenfassung	216
11	Zusammenfassung und Ausblick	219
	Referenzen	223
	Index	XIX

Abbildungsverzeichnis

2.1	Änderung der Bildinhalte bei unterschiedlichen Schnitten	10
2.2	Modellierung von weichen Schnitten	12
2.3	Schnitterkennung mit Hilfe von Orts-Zeit-Bildern	15
2.4	Erkennung von Schnitten mit Histogrammdifferenzen	16
2.5	Standardabweichung der Helligkeitswerte eines Bildes	18
2.6	Analyse der Kantenänderungsrate	20
2.7	Zusammenhang zwischen kumulierten Histogrammen und der Earth-Movers-Distanz	36
3.1	Schätzung der Bewegungsvektoren	43
3.2	Auswahl geeigneter Bewegungsvektoren	45
3.3	Transformation von Bildern	49
3.4	Änderung der Kameraparameter in einer Filmsequenz	51
4.1	Lineare Interpolation eines Pixels	58
4.2	Berechnung des Bildhintergrundes	60
4.3	Fehlerhafte Hintergrundbilder	61
4.4	Differenz zwischen transformierten Bildern	62
4.5	Morphologische Operatoren	64
4.6	Segmentierungsergebnisse	65
4.7	Automatisch segmentierte Objekte und Panoramabilder	68
4.8	Einfügen von Objekten in Videosequenzen	69
5.1	Kontur einer Person im Zeitablauf	73
5.2	Glättung einer Kontur	77
5.3	Abbildung im krümmungsbasierten Skalenraum	78

5.4	Bögen konvexer Regionen im Skalenraumbild	79
5.5	Auswirkung von Rauschen auf Skalenraumabbildungen	82
5.6	Mehrdeutigkeiten in Skalenraumabbildungen	83
5.7	Transformation einer Kontur	85
5.8	Ermittlung transformierter Konturpixel	86
5.9	Punkte innerhalb und außerhalb von Konturen	87
5.10	Beispielobjekte der Datenbank	90
5.11	Klassifikationsergebnisse	96
5.12	Beispiele für nicht erkannte Objekte	100
5.13	Objekterkennung in historischen Videos	102
6.1	Horizontales Projektionsprofil	110
6.2	Erkennung der Textzeilen eines Bildes	111
6.3	Buchstabengrenzen innerhalb einer Textzeile	112
6.4	Optimierung des Kürzeste-Pfade-Algorithmus	113
6.5	Segmentierung der Textpixel	115
6.6	Merkmale zur Charakterisierung von Buchstaben	116
6.7	Beispiele für Skalenraumabbildungen	117
6.8	Buchstaben der Datenbank	118
6.9	Beispiele für verrauschte Buchstaben	120
6.10	Ergebnisse der Texterkennung	123
7.1	Klassifikation von Algorithmen zur Gesichtserkennung	128
7.2	Struktur eines neuronalen Netzes	132
7.3	Erkennung von Gesichtsregionen	135
7.4	Erkennung von Gesichtsmerkmalen	136
7.5	Normierung eines Gesichtes	137
7.6	Beispiele für Eigengesichter	138
7.7	Anordnung der Gesichter in einem Video	142
8.1	Klassifikation der Verfahren zur Adaption von Videos	147
8.2	Adaption der Farbtiefe	152
8.3	Transformation eines Farbbildes in ein Binärbild	154
8.4	Adaption der Bildauflösung eines Videos	157
8.5	Beispiele für die semantische Adaption eines Videos	159

8.6	Experimentelle Ergebnisse zur Adaption der Farbtiefe	169
8.7	Experimentelle Ergebnisse zur Adaption der Bildauflösung	170
8.8	Experimentelle Ergebnisse zur Adaption historischer Videos	171
9.1	Erzeugung computergenerierter Zusammenfassungen	177
9.2	Systemüberblick	183
9.3	Gruppierung ähnlicher Kameraeinstellungen	186
9.4	Auswahl von Kameraeinstellungen	189
9.5	Maß zur Beurteilung der Verteilung der Kameraeinstellungen	194
9.6	Bewertung von Kameraeinstellungen	195
9.7	Beispiele einer statischen Zusammenfassung	199
9.8	Statische Zusammenfassungen in Form einer Kollage	200
9.9	Ergebnisse der Evaluation	203
10.1	Analyse der Objekt- und Personenbewegungen	209
10.2	Ermittlung der Objektklasse	212
10.3	Ergebnisse zur Analyse der Fahrt eines PKWs	213
10.4	Ergebnisse zur Bewegungsanalyse von Personen	217

Tabellenverzeichnis

2.1	Klassifikation eines Schnittes	9
2.2	Auswirkung der Anzahl schwacher und starker Kanten auf den kantenbasier- ten Kontrast	22
2.3	Verteilung der Schnitte in den ausgewählten Videosequenzen	23
2.4	Theoretische Obergrenzen für die Erkennung harter Schnitte.	25
2.5	Theoretische Obergrenzen der Klassifikationsergebnisse für harte Schnitte mit optimierten Verfahren	26
2.6	Optimale Schwellwerte für harte Schnitte	27
2.7	Klassifikationsergebnisse für Ein-, Aus- und Überblendungen	30
2.8	Optimale Parameter für weiche Schnitte	31
2.9	Klassifikationsergebnisse für harte und weiche Schnitte	32
2.10	Klassifikationsergebnisse für harte Schnitte in historischen Videos	34
3.1	Zusammenhang zwischen Kameraoperation und den Parametern des Kamera- modells	48
3.2	Gültige Parameter des Kameramodells	50
3.3	Klassifikationsergebnisse für das Kameramodell	52
3.4	Automatisch erkannte Kameraoperationen der Testsequenzen	53
4.1	Testsequenzen zur automatischen Objektsegmentierung	66
5.1	Objekte und Objektklassen der Datenbank	91
5.2	Klassifikationsergebnisse zur Objekterkennung	93
5.3	Anwendung globaler Konturdeskriptoren	97
5.4	Klassifikationsergebnisse zur Objekterkennung mit optimierten Verfahren . .	98
6.1	Erkennungsraten bei unterschiedlichen Zeichensätzen	119

6.2	Ergebnisse zur Segmentierung der Buchstaben	121
6.3	Ergebnisse zur Klassifikation der Buchstaben	122
7.1	Ergebnisse der Gesichtserkennung	140
9.1	Merkmale zur Beschreibung von Kameraeinstellungen	190
10.1	Objektklassen und Unterklassen der Datenbank	210
10.2	Anteil der fehlerhaft klassifizierten Objekte und Personen	214

KAPITEL 1

Einleitung

Der Übergang von analogen zu digitalen Videos hat in den letzten Jahren zu großen Veränderungen innerhalb der Filmarchive geführt. Durch die Digitalisierung der Filme ergeben sich für Archive neue Möglichkeiten. Die Auswirkungen des Wechsels von analogen Filmrollen zu digital gespeicherten Videos sind langfristig nur schwer abschätzbar. Für digitale Videos sollte gewährleistet sein, dass sie auf zukünftiger Hard- und Software wiedergegeben werden können.

Andererseits bieten digitale Videos deutliche Vorteile gegenüber analogen Filmen. Eine Abnutzung oder Alterung der Filmrollen ist ausgeschlossen, so dass die Qualität unverändert erhalten bleibt. Zudem wird ein netzbasierter und somit deutlich einfacherer Zugriff auf die Videos in den Archiven möglich. Zusätzliche Dienste stehen den Archivaren und Anwendern zur Verfügung, die erweiterte Suchmöglichkeiten bereitstellen und die Navigation bei der Wiedergabe erleichtern. Die Suche innerhalb der Videoarchive erfolgt mit Hilfe von Metadaten, die weitere Informationen über die Videos zur Verfügung stellen. Ein großer Teil der Metadaten wird manuell von Archivaren eingegeben, was mit einem großen Zeitaufwand und hohen Kosten verbunden ist.

Durch die computergestützte Analyse eines digitalen Videos ist es möglich, den Aufwand bei der Erzeugung von Metadaten für Videoarchive zu reduzieren. In dieser Arbeit werden neue Verfahren vorgestellt, um wichtige semantische Inhalte der Videos zu erkennen. Unter dem Begriff *Semantik* wird im Folgenden der visuelle Inhalt verstanden, der in Bildern, Bildsequenzen und Videos dargestellt ist. Algorithmen zur semantischen Analyse, auf die in

dieser Arbeit eingegangen wird, ermitteln beispielsweise alle Personen innerhalb einer Kameraeinstellung oder erkennen die Art der Bewegung einer Person. Tiefergehende semantische Inhalte, wie beispielsweise die Frage, warum sich eine Person in bestimmter Weise verhält, können mit dem heutigen Stand der Forschung nicht beantwortet werden. Dennoch sind die automatisch ermittelten semantischen Informationen sehr wertvoll, da sie die Arbeit mit digitalen Videoarchiven erleichtern. Die Informationen unterstützen nicht nur die Suche in den Archiven, sondern führen auch zur Entwicklung neuer Anwendungen. Beispielsweise können computergenerierte Zusammenfassungen von Videos erzeugt oder Videos automatisch an die Eigenschaften des Abspielgerätes angepasst werden.

Im Rahmen des Projektes *European Chronicles Online*¹ wurde eine komplexe Anwendung entwickelt, um Archive mit historischen Videos zu verwalten und die historisch wertvollen Filme den Archivaren und der Öffentlichkeit leichter zugänglich zu machen. Die im Archiv gespeicherten historischen Filme wurden von vier europäischen Filmarchiven für das Projekt zur Verfügung gestellt. Ein großer Teil der in dieser Arbeit entwickelten Algorithmen sind in das *European-Chronicles-Online*-System integriert. Die im *European-Chronicles-Online*-Archiv gespeicherten Filme wurden Anfang bis Mitte des letzten Jahrhunderts gedreht und in den letzten Jahren digitalisiert. Durch die Lagerung und Abnutzung der Filmrollen über mehrere Jahrzehnte sind viele Videos stark verrauscht und enthalten deutlich sichtbare Bildfehler. Die Bildqualität der historischen Schwarz-Weiß-Filme unterscheidet sich signifikant von der Qualität aktueller Videos, so dass eine verlässliche Analyse mit bestehenden Verfahren häufig nicht möglich ist. Im Rahmen dieser Arbeit werden neue Algorithmen vorgestellt, um eine zuverlässige Erkennung von semantischen Inhalten auch in historischen Videos zu ermöglichen. Die Arbeit ist in zwei Teile untergliedert. Im ersten Teil werden Algorithmen zur automatischen Analyse struktureller und semantischer Inhalte eines Videos vorgestellt. Die Anwendungen des zweiten Teils nutzen die computergenerierten Inhalte der Analysealgorithmen. Da sich die Verfahren der einzelnen Kapitel thematisch deutlich voneinander unterscheiden, werden Vorarbeiten und Ergebnisse innerhalb der einzelnen Kapitel vorgestellt. Zentrale Bestandteile dieser Arbeit sind die Kapitel zur Objekterkennung, Adaption von Videos und automatischen Erzeugung von Zusammenfassungen, in denen wesentliche neue Ideen vorgestellt werden.

Im zweiten Kapitel des ersten Teils werden Algorithmen zur *Schnitterkennung* und zur Identifikation der einzelnen Kameraeinstellungen betrachtet. Die Erkennung harter und weicher

¹Auf das *European-Chronicles-Online*-Projekt wird näher im Rahmen der Schnitterkennung von historischen Videodokumentationen eingegangen.

Schnitte ist Voraussetzung für die weiteren Analyseschritte, da sich semantische Inhalte eines Videos häufig auf Kameraeinstellungen beziehen. Wir haben neue optimierte Verfahren entwickelt, die insbesondere für eine zuverlässige Schnitterkennung bei historischen Videos erforderlich sind.

In Kapitel 3 werden Algorithmen zur Berechnung der *Kamerabewegung* erläutert. Die Veränderung der Kamerabewegung zwischen zwei aufeinander folgenden Bildern wird durch ein Modell beschrieben. Die Identifikation von Kameraschwenks, Zoomoperationen und verwackelten Kameraeinstellungen erfolgt durch Analyse der Parameter des Kameramodells. Im Rahmen der experimentellen Ergebnisse wird speziell darauf eingegangen, wie fehlerhafte Parameter des Kameramodells identifiziert werden können.

Die Kamerabewegung wird für die *bewegungsbasierte Segmentierung von Objekten* benötigt, auf die in Kapitel 4 näher eingegangen wird. Durch einen Ausgleich der Kamerabewegung werden Hintergrundbilder erzeugt, in denen Vordergrundobjekte nicht mehr enthalten sind. Durch einen Vergleich mit dem Hintergrundbild werden alle Objekte, deren Positionen sich im Zeitablauf verändern, segmentiert.

Kapitel 5 ist eines der zentralen Kapitel dieser Arbeit, in dem wesentliche neue Ideen und Algorithmen zur *Erkennung von Objekten* vorgestellt werden. Mit Hilfe von Skalenraumabbildungen werden Merkmale der äußeren Kontur eines Objektes abgeleitet und mit Merkmalen bekannter Konturen verglichen. Wir haben zwei neue Algorithmen entwickelt, durch die Mehrdeutigkeiten in den Skalenraumabbildungen vermieden und konvexe Objektregionen beim Konturvergleich berücksichtigt werden. Zusätzlich wird ein neues Verfahren zur Aggregation der Klassifikationsergebnisse für Videosequenzen vorgestellt. Die Objekterkennungsalgorithmen sind Bestandteil des *European-Chronicles-Online-Systems*, in dem für jedes Video Informationen über Objekte automatisch zur Verfügung gestellt werden.

Verfahren zur *Erkennung von Textregionen und Buchstaben* werden in Kapitel 6 eingeführt. Im Vergleich zu eingescannten Dokumenten stellt die Segmentierung eines Textes wegen des häufig komplexen Bildhintergrundes und der geringen Bildauflösung des Videos eine besondere Herausforderung dar. Neue Algorithmen werden erläutert, um Trenner zwischen Buchstaben zu identifizieren und eine zuverlässige Segmentierung der einzelnen Buchstaben zu ermöglichen.

Im siebten und letzten Kapitel des ersten Teils werden Algorithmen zur *Gesichtserkennung* vorgestellt. Die Klassifikation erfolgt in einem dreistufigen Verfahren. Nach der Lokalisierung der Gesichtsregionen folgt die Segmentierung eines Gesichtes, bei der Skalierungsunterschiede, Rotationen, Beleuchtungsunterschiede und der Kontrast ausgeglichen werden. Im letzten

Schritt erkennt der Algorithmus die normierten Gesichter.

Der zweite Teil dieser Arbeit beschreibt interessante neue Anwendungen, welche die Ergebnisse der semantischen Analyse eines Videos nutzen. Zuerst werden Verfahren zur *Adaption von Videos* in Kapitel 8 betrachtet. Die Adaptionalgorithmen ermöglichen eine automatische Anpassung eines Videos an die unterschiedlichen Eigenschaften von Anzeigegeräten. Neue Algorithmen und Ideen werden zur Adaption der Farbtiefe, Anpassung der Bildauflösung und Verbesserung der Bildqualität entwickelt.

In Kapitel 9 werden semantische Inhalte eines Videos identifiziert, um *automatische Zusammenfassungen von Videos* zu erzeugen. Eine Zusammenfassung kann als Sammlung von aussagekräftigen Bildern oder als kurze Videosequenz gespeichert werden. Neue Heuristiken zur Auswahl und Kombination der Bilder bzw. Kameraeinstellungen werden eingesetzt, um die wesentlichen semantischen Inhalte des Videos zu erhalten.

Eine Anwendung zur *Analyse der Bewegungen von Objekten und Personen* wird in Kapitel 10 vorgestellt. Insbesondere durch die Analyse der Änderungen eines Objektes im Zeitablauf können detaillierte Informationen beispielsweise über die Fahrt eines PKWs oder die Bewegungsabläufe einer Person ermittelt werden.

Die Arbeit wird mit einer Zusammenfassung und einem Ausblick abgeschlossen.

Teil I

Algorithmen zur automatischen Analyse von Videos

KAPITEL 2

Erkennung von Schnitten in Videos

Die Schnitterkennung ist ein zentraler erster Schritt bei der computergestützten Analyse eines Videos. In diesem Kapitel werden Algorithmen zur Erkennung der unterschiedlichen Arten von Schnitten in Videos vorgestellt und analysiert. Da viele unterschiedliche Verfahren zur Erkennung von Schnitten in den letzten Jahren entwickelt wurden, sollen in diesem Kapitel nur einzelne ausgewählte Verfahren vorgestellt und detailliert analysiert werden. Des Weiteren führt dieses Kapitel wesentliche für diese Arbeit grundlegende Begriffe und Verfahren ein.

Zur Analyse der Schnitterkennungsverfahren werden sowohl aktuelle Videos aus unterschiedlichen Genres betrachtet als auch historische Schwarz-Weiß-Videodokumentationen verwendet. Dabei werden zunächst optimale Schwellwerte für eine Gruppe von Testvideos ermittelt, anhand derer theoretische Obergrenzen für die einzelnen Schnitterkennungsverfahren abgeleitet werden. Anschließend wird mit Hilfe einer zweiten Gruppe von Testvideos überprüft, wie zuverlässig die Schnitterkennungsergebnisse mit den zuvor ermittelten Schwellwerten sind. Nach der Analyse aktueller Videos wird speziell auf die Schnitterkennung in historischen Videodokumentationen eingegangen. Rauschen und Bildfehler der Schwarz-Weiß-Filme führen zu wesentlich höheren Fehlerraten, so dass neue Metriken und Verfahren erforderlich sind, um gute Klassifikationsergebnisse auch in historischen Filmen zu erhalten.

Schnitte liefern Informationen über den Produktionsprozess eines Filmes, bei dem zunächst Rohmaterial erzeugt und in einem zweiten Schritt zu dem eigentlichen Film zusammengeschnitten wird. *Schnitte* (engl. *cut*) trennen kontinuierliche Aufnahmen, die als *Kameraeinstellungen* (engl. *shot*) bezeichnet werden. Die englische Berufsbezeichnung *cutter* stammt noch aus der Zeit, als Filme ausschließlich manuell geschnitten und neu zusammengefügt wurden.

Heute erfolgt die Bearbeitung des Rohmaterials überwiegend am Rechner. Die durchschnittliche Länge einer Kameraeinstellung der im Rahmen der experimentellen Ergebnisse analysierten aktuellen Videos und der historischen Videodokumentationen liegt bei weniger als 5 Sekunden. Wegen der geringen durchschnittlichen Länge soll im Rahmen dieser Arbeit für die Analyseschritte der folgenden Kapitel eine Kameraeinstellung, obwohl sie aus vielen *Einzelbildern* (engl. *frame*) besteht, als kleinste Einheit eines Filmes interpretiert werden, bei der die zeitliche Dimension noch enthalten ist.

Inhaltlich ähnliche und zeitlich aufeinanderfolgende Kameraeinstellungen werden als *Szenen* (engl. *scene*) bezeichnet. *Dialoge* sind spezielle Szenen, bei denen das Bild wiederholt zwischen zwei oder mehreren Personen wechselt. Die Informationen über Kameraeinstellungen dienen als Grundlage für nachfolgende Analyseschritte von Videos und sind Voraussetzung für die in den folgenden Kapiteln vorgestellten Verfahren zur Objekt-, Gesichts- oder Texterkennung.

In diesem Kapitel werden in Abschnitt 2.1 zunächst die unterschiedlichen Arten von Schnitten vorgestellt. In Abschnitt 2.2 folgt eine Beschreibung der Algorithmen zur automatischen Schnitterkennung, wobei zunächst eine Modellierung der unterschiedlichen Schnitteffekte erfolgt. Es wird insbesondere auf grundlegende Verfahren zum Vergleich von Bildern eingegangen und erläutert, welche Ähnlichkeitsmaße zur Erkennung harter und weicher Schnitte geeignet sind. In Abschnitt 2.3 werden experimentelle Ergebnisse für Videos aus unterschiedlichen Genres betrachtet, verbesserte Verfahren zur Schnitterkennung vorgeschlagen und Ursachen für Klassifikationsfehler analysiert. Zum Abschluss des Kapitels wird das Projekt *European Chronicles Online* vorgestellt, in dem ein komplexes System zur Verwaltung von Archiven mit historischen Videos in Zusammenarbeit mit mehreren Partnern entstanden ist. Da die Bildqualität der historischen Schwarz-Weiß-Filme mit der Qualität aktueller Filme nicht vergleichbar ist, führen bestehende Schnitterkennungsverfahren zu sehr schlechten Ergebnissen. Neue von uns entwickelte und in das *European-Chronicles-Online*-System integrierte Algorithmen zur Schnitterkennung für historische Filme werden abschließend vorgestellt.

2.1 Klassifikation eines Schnittes

Der Übergang von einer Kameraeinstellung zur folgenden wird als harter oder weicher Schnitt bezeichnet. In *harten Schnitten* (engl. *hard cut*) gibt es keinen Übergang zwischen den beiden Kameraeinstellungen. Im Falle eines *weichen Schnittes* (engl. *soft cut*) wird ein künstlicher Übergang zwischen den beiden Kameraeinstellungen erzeugt [185]. Innerhalb der analysier-

	Dauer eines Schnittes	
	Schnitt zwischen zwei Bildern	Schnitt über mehrere Bilder
Änderung der Werte (einzelner) Pixel zwischen benachbarten Bildern	Harter Schnitt	Wischeffekt
Kontinuierliche Änderung aller Pixelwerte über einen längeren Zeitraum		Ein- oder Ausblendung Überblendung

Tabelle 2.1: Klassifikation eines Schnittes

ten Videos treten am häufigsten harte Schnitte (92 Prozent) gefolgt von Überblendungen (6 Prozent) und Ein- oder Ausblendungen (1,9 Prozent) auf.

Bei einer *Überblendung* (engl. *dissolve*) erfolgt der Wechsel von einer Kameraeinstellung zur nächsten kontinuierlich. Zu Beginn der Überblendung sind die Bilder der ersten Kameraeinstellung vollständig sichtbar, die der Zweiten sind transparent. Im Verlauf der Überblendung nimmt die Transparenz der Bilder der ersten Kameraeinstellung zu und gleichzeitig die der Zweiten ab, so dass in den mittleren Bildern der Überblendung die Inhalte beider Kameraeinstellungen sichtbar sind.

Ein- und Ausblendungen (engl. *fade in, fade out*) sind spezielle Überblendungen, bei denen eine der beiden Kameraeinstellungen aus monochromen – häufig schwarzen – Bildern besteht. Wesentlich seltener ($< 0,1$ Prozent) werden in den analysierten Videos Wischeffekte (engl. *wipe*) verwendet. Statt die Intensität aller Pixelwerte kontinuierlich zu verändern, werden Pixel ausgewählter Bildregionen sofort verändert. Die Wischeffekte laufen häufig horizontal oder vertikal durch das Bild, so dass in den mittleren Bildern eines Wischeffektes in einer Bildhälfte die Inhalte der alten und in der anderen die der neuen Kameraeinstellung sichtbar sind.

Die Dauer eines weichen Schnittes variiert zwischen einem Bruchteil einer Sekunde und mehreren Sekunden. Eine Überblendung mit einer Länge eines einzelnen Bildes hält ein Betrachter für einen harten Schnitt, wobei der Übergang etwas weniger plötzlich empfunden wird. Überblendungen über einen Zeitraum von mehreren Sekunden werden von Regisseuren bewusst eingesetzt, um spezielle Wirkungen beim Zuschauer – wie z. B. den Beginn eines Traumes – zu erzielen. Tabelle 2.1 veranschaulicht die am häufigsten auftretenden Schnitte in Videos. Schnitte lassen sich nach ihrer Dauer und der Art der Änderung der Pixelwerte klassifizieren. Abbildung 2.1 zeigt Beispiele für die Änderung der Bildinhalte im Zeitablauf in Abhängigkeit der unterschiedlichen Schnitte.

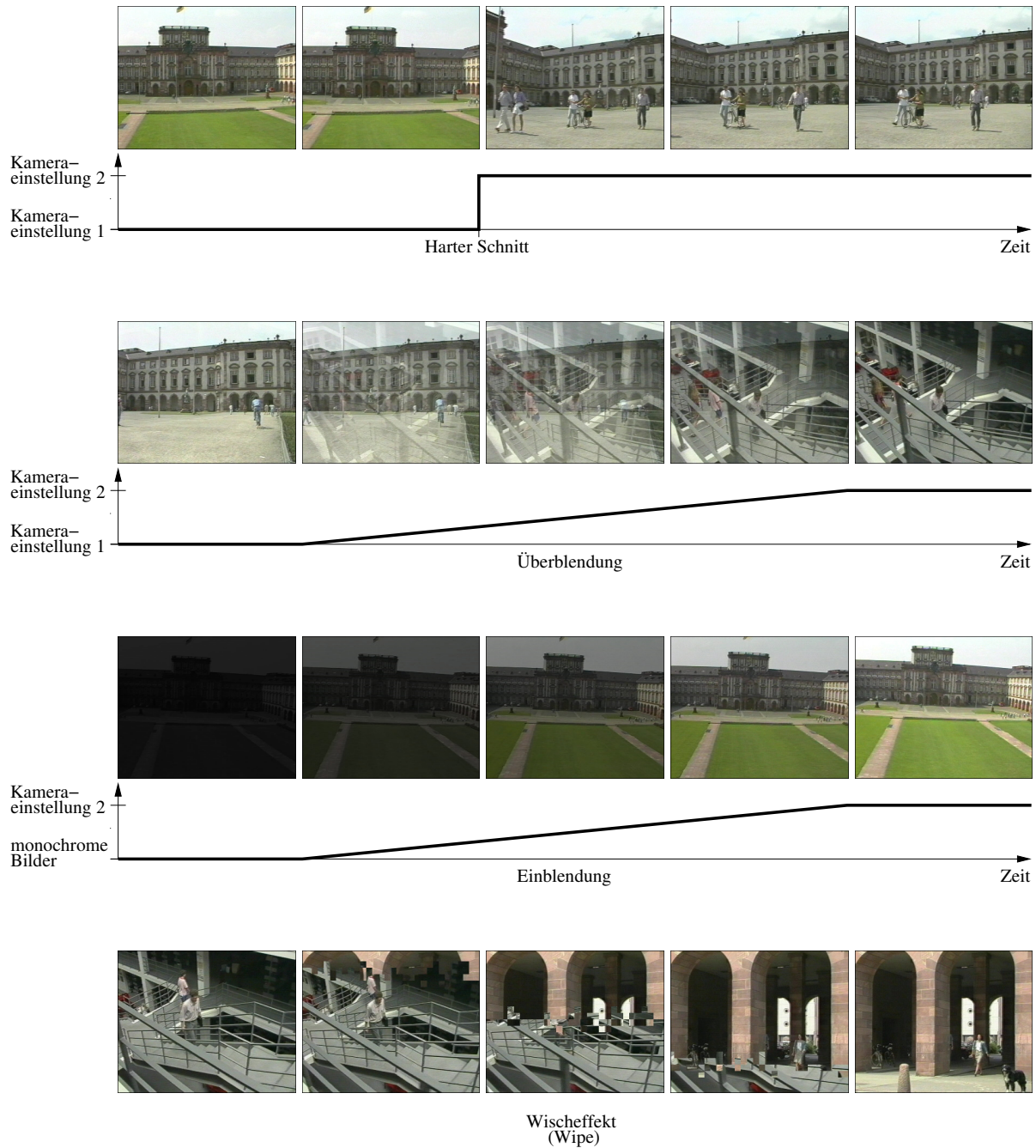


Abbildung 2.1: Änderung der Bildinhalte in Abhängigkeit eines Schnittes

2.2 Computergestützte Erkennung eines Schnittes

Für die automatische Erkennung von harten Schnitten wird die Ähnlichkeit zweier Bilder I_i und I_j mit $1 \leq i < j \leq N$ innerhalb einer Videosequenz $(I_1 \dots I_N)$ analysiert. Eine Kameraeinstellung wird durch ein zusammenhängendes zeitliches Intervall innerhalb des Videos spezifiziert. Es wird die grundlegende Annahme getroffen, dass die Unterschiede der Bilder innerhalb einer Kameraeinstellung wesentlich geringer sind als die Unterschiede von Bildern unterschiedlicher Kameraeinstellungen.

Die Erkennung harter und weicher Schnitte lässt sich als dreistufiges Verfahren abbilden [44]: In einem ersten Schritt wird eine geeignete Abbildung τ definiert, die ein Bild in einen Merkmalsraum transformiert. Mit Hilfe eines robusten Distanzmaßes D wird anschließend die Ähnlichkeit zweier Bilder anhand ihrer Merkmalswerte bestimmt. Beim dritten Schritt geht die Annahme ein, dass die Merkmalswerte der Bilder innerhalb einer Kameraeinstellung geringere Unterschiede aufweisen als Bilder unterschiedlicher Kameraeinstellungen. Dazu wird ein geeigneter Schwellwert T festgelegt und die Distanz zweier Bilder mit diesem Wert verglichen. Beim Überschreiten des Wertes wird angenommen, dass ein Schnitt zwischen den beiden Bildern vorliegt [580, 584]. Bewegungen und Helligkeitsänderungen können auch innerhalb von Kameraeinstellungen deutliche Distanzwerte verursachen. Falls statt eines *absoluten Schwellwertes* ein *adaptiver Schwellwert* verwendet wird, sind in der Regel zuverlässigere Klassifikationsergebnisse möglich.

Die Transformation der Bilder I_i einer Videosequenz in einen Merkmalsraum wird definiert als

$$\tau : \mathbb{N}^m \rightarrow F, \quad (2.1)$$

wobei \mathbb{N}^m den Raum definiert, der durch alle Bilder $(I_n \in \mathbb{N}^m)$ aufgespannt wird. F spezifiziert den Merkmalsraum mit $\tau(I_n) \in F$. Das Distanzmaß D beurteilt auf Basis der Merkmalswerte die Unterschiede zwischen zwei Bildern:

$$D : F \times F \rightarrow \mathbb{R}^+. \quad (2.2)$$

Dabei soll die Distanz $D_{i,j}$ ein Maß für die visuelle Ähnlichkeit zweier Bilder i und j liefern. Es wird angenommen, dass ein Schnitt zwischen den Bildern i und j vorliegt, falls gilt:

$$D_{i,j} = D(\tau(I_i), \tau(I_j)) > T_{i,j}. \quad (2.3)$$

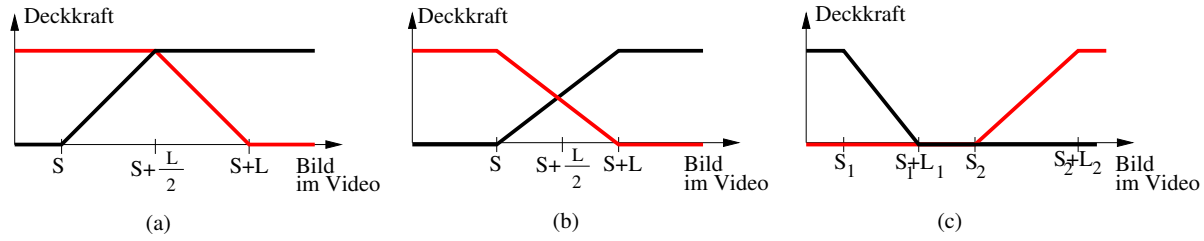


Abbildung 2.2: Modellierung von additiven Überblendungen (a), Kreuz-Überblendungen (b) und Aus- bzw. Einblendungen (c).

Bei dem Schwellwert $T_{i,j}$ muss es sich nicht um einen absoluten Wert handeln. Vielmehr kann $T_{i,j}$ auch als adaptiver Schwellwert abhängig von den Distanzen der zu i und j benachbarten Bilder festgelegt werden. Um zwischen harten und weichen Schnitten zu unterscheiden, wird der kleinste zeitliche Abstand ($j - i$) zwischen zwei Bildern ermittelt, bei dem ein Schnitt erkannt wird. Liegt der Schnitt zwischen zwei benachbarten Bildern ($j = i + 1$), so handelt es sich um einen harten Schnitt, ansonsten um einen weichen Schnitt.

Um eine Differenzierung der unterschiedlichen weichen Schnitte zu ermöglichen, werden die Eigenschaften von Überblendungen sowie Ein- und Ausblendungen näher betrachtet [68]. Bei einer *additiven Überblendung* (engl. *additive dissolve*) bleibt während des Einblendens der zweiten Kameraeinstellung die erste Kameraeinstellung sichtbar, und erst wenn die zweite Kameraeinstellung vollständig sichtbar ist, beginnt die Ausblendung der zweiten Kameraeinstellung. Bei einer *Kreuz-Überblendung* (engl. *cross dissolve*) erfolgt das Ausblenden der ersten Kameraeinstellung gleichzeitig mit dem Einblenden der zweiten Kameraeinstellung. Eine Ein- bzw. Ausblendung lässt sich durch eine Kreuz-Überblendung beschreiben, bei der eine der beiden Kameraeinstellungen monochrome Bilder enthält. Abbildung 2.2 verdeutlicht schematisch die Unterschiede den unterschiedlichen Arten der Überblendungen. Der Startzeitpunkt einer Überblendung wird mit S und die Dauer mit L bezeichnet.

Bei der Modellierung von Ein-, Aus- und Überblendungen wird im Folgenden angenommen, dass die Veränderung der Transparenz durch eine lineare Funktion approximiert werden kann. Falls sich der Bildinhalt beider Kameraeinstellungen nicht verändert, ist eine exakte Spezifikation des Bildinhaltes während einer Überblendung möglich:

$$I_k = \alpha_k \cdot I_S + \beta_k \cdot I_{S+L} \quad \text{mit} \quad S \leq k \leq S + L. \quad (2.4)$$

Für additive Überblendungen werden die Parameter α und β durch

$$\alpha_k = \begin{cases} 1 & \text{für } S \leq k \leq S + L/2 \\ 1 - \frac{k-S-L/2}{L/2} & \text{für } S + L/2 < k \leq S + L \end{cases} \quad (2.5)$$

$$\beta_k = \begin{cases} \frac{k-S}{L/2} & \text{für } S \leq k \leq S + L/2 \\ 1 & \text{für } S + L/2 < k \leq S + L, \end{cases} \quad (2.6)$$

definiert, für Kreuz-Überblendungen und Ausblendungen durch

$$\alpha_k = 1 - \frac{k-S}{L} \quad \text{und} \quad \beta_k = 1 - \alpha_k \quad \text{mit} \quad S \leq k \leq S + L. \quad (2.7)$$

Bei Ausblendungen ist es möglich, statt schwarzer Bilder auch $\beta_k = 0$ zu setzen. Einblendungen werden durch ein Vertauschen der Parameter α und β modelliert. Bei der Wahl eines geeigneten Distanzmaßes führt die lineare Veränderung der Transparenz während der Ein-, Aus- oder Überblendung zu einer gleichmäßigen Änderung der Distanzmaße zwischen jeweils zwei benachbarten Bildern:

$$D_{i,i+1} \approx D_{i+1,i+2} \quad \forall \quad S \leq i < S + \frac{L}{2} - 1 \quad \text{und} \quad S + \frac{L}{2} \leq i < S + L - 1. \quad (2.8)$$

Zusätzlich nimmt die Differenz mit steigendem Abstand zwischen den Bildern innerhalb eines weichen Schnittes zu:

$$D_{i,i+j} < D_{i,i+k} \quad \forall \quad S \leq i < i+j < i+k \leq S + L. \quad (2.9)$$

Falls ein Schnitt zwischen zwei *nicht* benachbarten Bildern i und j identifiziert wurde, muss zunächst anhand der Gleichungen 2.8 und 2.9 überprüft werden, ob Ein-, Aus- oder Überblendungen modelliert werden können. In den im Rahmen der experimentellen Ergebnisse analysierten Videosequenzen treten weitere Arten von Schnitten wie beispielsweise Wischeffekte nur sehr vereinzelt auf, so dass eine detailliertere Analyse zusätzlicher Schnitteffekte nicht vorgenommen wird.

2.2.1 Pixelbasierte Verfahren zur Schnitterkennung

Die Summe der absoluten Pixeldifferenzen D_{SAD} der beiden Bilder I_i und I_j ist ein einfach zu berechnendes Distanzmaß zur Erkennung harter Schnitte:

$$D_{SAD} = \frac{1}{N_x \cdot N_y} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} |I_i(x, y) - I_j(x, y)|. \quad (2.10)$$

Das Distanzmaß wird mit der Bildgröße $N_x \cdot N_y$ normiert. Ein wesentlicher Vorteil dieses Distanzmaßes besteht darin, dass der Bildraum \mathbb{N}^m mit dem Merkmalsraum F identisch ist und eine Abbildung $\tau(I_n)$ vom Bildraum in den Merkmalsraum nicht erforderlich ist. Es gelten zudem die Bedingungen der Gleichungen 2.8 und 2.9, so dass während einer Überblendung die Differenzen benachbarter Bilder ähnliche Werte annehmen und mit zunehmender zeitlicher Distanz zwischen zwei Bildern die Differenzen ansteigen [68, 69]. Auch zur Erkennung der Art eines Wischeffektes eignet sich die Summe der absoluten Differenzen, indem ein Binärbild erzeugt wird, in dem signifikante Pixeldifferenzen markiert sind. Die Analyse der Position und Bewegungsrichtung des Schwerpunktes der Pixel im Differenzbild ermöglicht die Erkennung und Beschreibung eines Wischeffektes.

Drew und Ngo erzeugen *Orts-Zeit-Bilder* aus Videos [122, 382, 384]. Als charakteristisches Merkmal wird aus jedem Bild im Video die mittlere Pixelzeile oder Pixelspalte ausgewählt und bildet eine Zeile im Orts-Zeit-Bild. Das Distanzmaß bildet spezifische Strukturen im Orts-Zeit-Bild ab und ermöglicht die Erkennung von harten und weichen Schnitten. Harte Schnitte zeigen waagrechte Änderungen im Bild, wohingegen Wischeffekte eine diagonale Orientierung aufweisen. Abbildung 2.3 verdeutlicht die Erzeugung von Orts-Zeit-Bildern und die typischen Merkmale der unterschiedlichen Schnitte.

Alle auf Pixeldifferenzen basierenden Verfahren haben den Nachteil, dass hohe Fehlerraten bei Objekt- und Kamerabewegungen auftreten. Eine Person, die sich beispielsweise von links nach rechts durch ein Bild bewegt, erzeugt Änderungen der Pixeldifferenzen zwischen benachbarten Bildern, die mit einem horizontalen Wischeffekt vergleichbar sind.

2.2.2 Schnitterkennung mit Histogrammen

Histogrammbasierte Verfahren liefern bei geringer Komplexität gute Ergebnisse für die Erkennung harter Schnitte und werden in vielen Ansätzen verwendet [57, 65, 196, 340, 493]. Ein *Histogramm* speichert für jeden Grau- bzw. Farbwert die absolute oder relative Anzahl der Pixel dieser Helligkeit bzw. Farbe im Bild. Aussagen über die durchschnittliche Helligkeit, den Kontrast und die Farben eines Bildes lassen sich aus Histogrammen ableiten, die Anordnung der Farben im Bild jedoch nicht [187, 238, 354].

In 8-Bit-Graustufenbildern ist die Größe der Histogramme auf 256 Elemente beschränkt, wo-

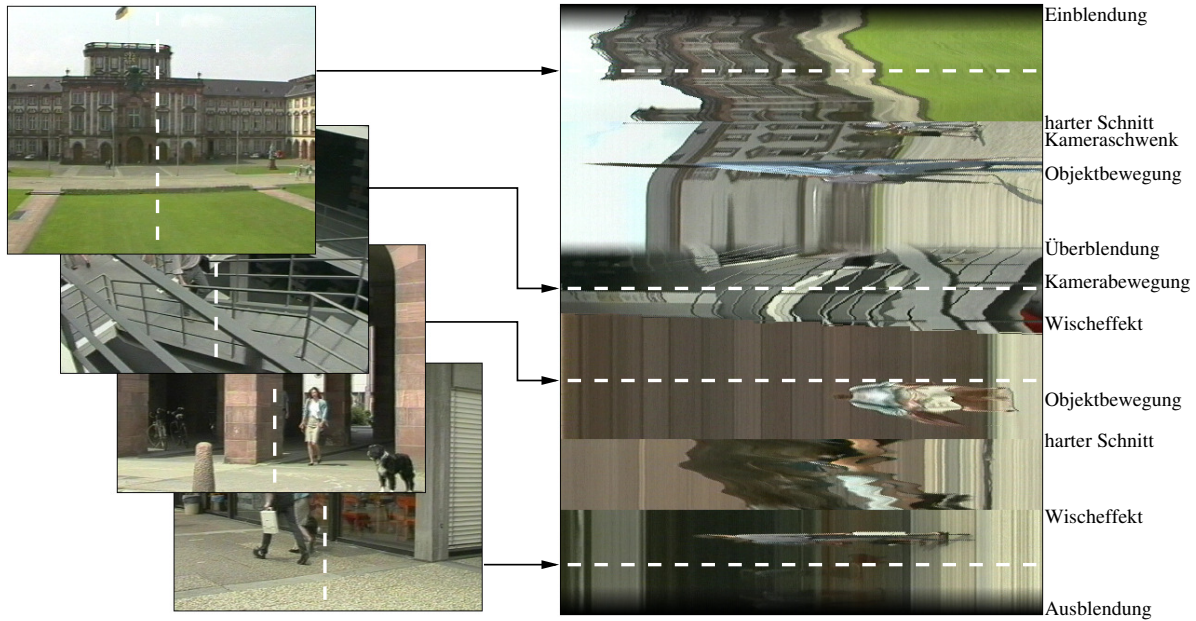


Abbildung 2.3: Links: Ausgewählte Bilder einer Videosequenz. Rechts: Im zugehörigen Orts-Zeit-Bild sind Schnitte sowie Objekt- und Kamerabewegungen markiert.

hingegen Farbbilder mit 24-Bit Farbtiefe theoretisch mehr als 16 Millionen unterschiedliche Farben enthalten können. Da Histogramme dieser Größenordnung nicht mehr aussagekräftig sind, wird zur Verringerung der Dimension des Merkmalsraumes F entweder die Anzahl der Farben reduziert, oder es werden für jeden Farbkanal getrennte Histogramme berechnet.

Mit einer Vielzahl unterschiedlicher Metriken lassen sich *Histogrammdifferenzen* berechnen [440]. Die *Minkowski-Metrik* L_p vergleicht die Elemente in zwei Histogrammen H_1 und H_2 und ist definiert als:

$$L_p(H_1, H_2) = \left(\sum_{m=1}^M |H_1(m) - H_2(m)|^p \right)^{\frac{1}{p}}. \quad (2.11)$$

M spezifiziert die Größe des Histogramms und p definiert die Norm der Metrik. Bei der Berechnung der Bilddifferenzen mit Hilfe von Histogrammen wird im Allgemeinen die L_1 - oder L_2 -Norm verwendet. Die L_1 -Norm (Summe der *absoluten* Histogrammdifferenzen) gewichtet kleine Differenzwerte stärker als die L_2 - oder *Euklidische Norm* (Summe der *quadrierten* Histogrammdifferenzen).

Die Erkennung harter Schnitte ist durch einen Vergleich der Histogrammdifferenzen benachbarter Bilder entsprechend der Gleichung 2.3 möglich. In Abbildung 2.4 sind Histogramm-

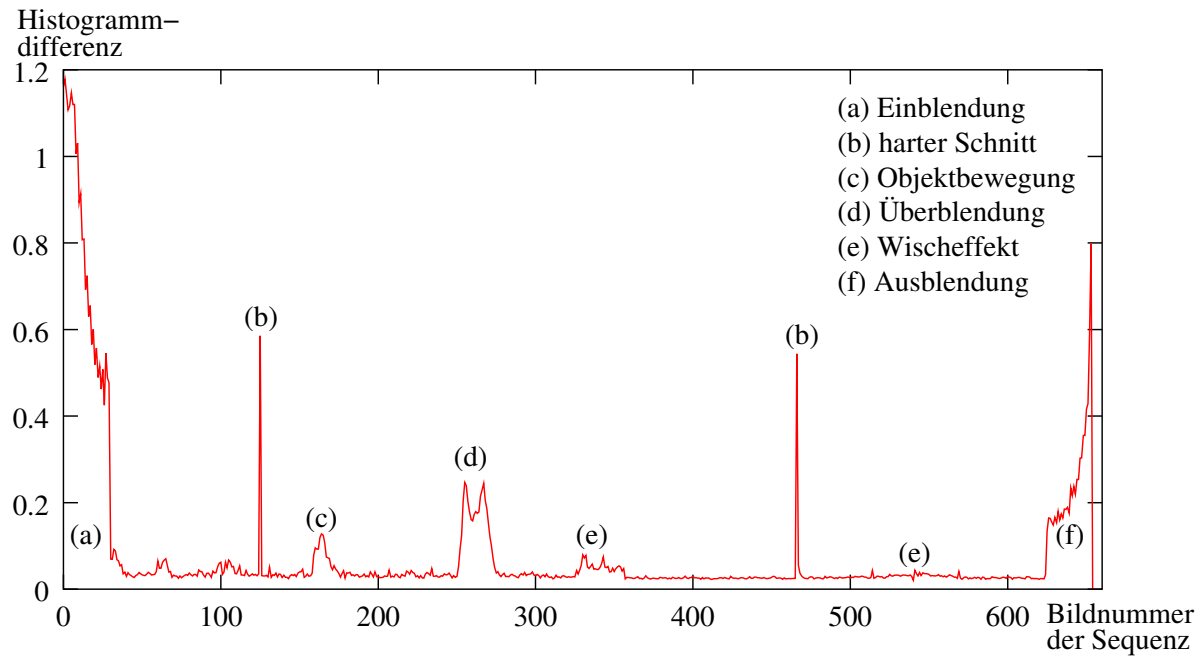


Abbildung 2.4: Histogrammdifferenzen benachbarter Bilder nach der L_1 -Norm in einer Videosequenz mit unterschiedlichen Schnitten

differenzen der L_1 -Norm einer Videosequenz mit mehreren Schnitten abgebildet. Für die Beispielsequenz in der Abbildung ist deutlich zu sehen, dass sich Histogrammdifferenzen zur Erkennung harter Schnitte gut eignen. Je nach Stärke einer Ein- oder Ausblendung können deutliche Histogrammdifferenzen zu Beginn einer Einblendung bzw. am Ende einer Ausblendung auftreten. Ohne Berücksichtigung weiterer für Ein- oder Ausblendungen charakteristischer Merkmale sind fehlerhafte Klassifikationen beim Vergleich der Histogrammdifferenzen benachbarter Bilder zu erwarten.

Die Histogrammdifferenzen zwischen benachbarten Bildern einer Videosequenz sind bei Überblendungen oder Wischeffekten häufig so gering, dass kein Unterschied zwischen Objektbewegungen und weichen Schnitten erkannt werden kann. Die Analyse der Histogrammdifferenzen $L_p(H_i, H_j)$ nicht benachbarter Bilder ($j - i > 1$) und der Vergleich innerhalb einer Überblendung benachbarter Bilder ermöglichen eine Erkennung weicher Schnitte [213]. Ein Nachteil ist die hohe Anzahl an fehlerhaft erkannten Schnitten, da auch Objekt- oder Kamerabewegungen über einen längeren Zeitraum den Bildinhalt und somit die Histogramme signifikant verändern können.

Ähnlich den pixelbasierten Distanzmaßen treten Fehlklassifikationen insbesondere bei plötzlichen Helligkeitsänderungen und schnellen Bewegungen großer Objekte auf. Zur Klassifikation

harter Schnitte sind Histogramme dennoch gut geeignet, da die Wahrscheinlichkeit relativ gering ist, dass die Farbverteilung von Bildern unterschiedlicher Kameraeinstellungen ähnlich ist, so dass nur wenige Schnitte nicht oder falsch erkannt werden.

2.2.3 Schnitterkennung durch Analyse der Standardabweichung

Auch deutlich stärker aggregierte Bilddaten können zur Erkennung von Schnitten herangezogen werden [242, 260, 318]. Die Analyse der Standardabweichung σ_I der Helligkeitswerte aller Pixel eines Bildes I ermöglicht die Erkennung von Ein-, Aus- und Überblendungen:

$$\sigma_I = \sqrt{\frac{1}{N_x \cdot N_y} \sum_{x=1}^{N_x} \sum_{y=1}^{N_y} (I(x, y) - \bar{I})^2}. \quad (2.12)$$

Die durchschnittliche Helligkeit der Pixel eines Bildes wird mit \bar{I} , die Helligkeit an der Bildposition (x, y) mit $I(x, y)$ bezeichnet. Die Standardabweichung sinkt an den Rändern von Ein- bzw. Ausblendungen deutlich. In der Mitte einer Überblendung sinkt die Standardabweichung geringfügig, da die Pixel in diesen Bereichen durchschnittliche Helligkeits- bzw. Farbwerte annehmen. Nach Glättung der Standardabweichung mit einem Gaußfilter [73, 247] können Überblendungen und Ein- oder Ausblendungen durch Suche lokaler Minima erkannt werden. Abbildung 2.5 zeigt die geglättete Standardabweichung der Helligkeitspixel einer Videosequenz mit unterschiedlichen Schnitten. Besonders auffällig sind die Minima im Zentrum einer Überblendung und die geringen Werte bei Ein- und Ausblendungen.

Um ein lokales Minimum innerhalb der geglätteten Standardabweichung zu ermitteln, wird beim Distanzmaß $D_{i,j}$ zwischen Ein- und Ausblendungen bzw. zwischen der ersten und der zweiten Hälfte einer Überblendung unterschieden. Im Falle einer Ausblendung bzw. der ersten Hälfte einer Überblendung wird der Kontrast der einzelnen Bilder mit dem Kontrast des letzten Bildes verglichen und die Differenzen aufsummiert:

$$D_{i,j} = \sum_{k=i}^{j-1} \max(\sigma_k - \sigma_j, 0). \quad (2.13)$$

Bei einer Einblendung und dem zweiten Teil einer Überblendung erfolgt der Vergleich mit dem ersten Bild:

$$D_{i,j} = \sum_{k=i+1}^j \max(\sigma_k - \sigma_i, 0). \quad (2.14)$$

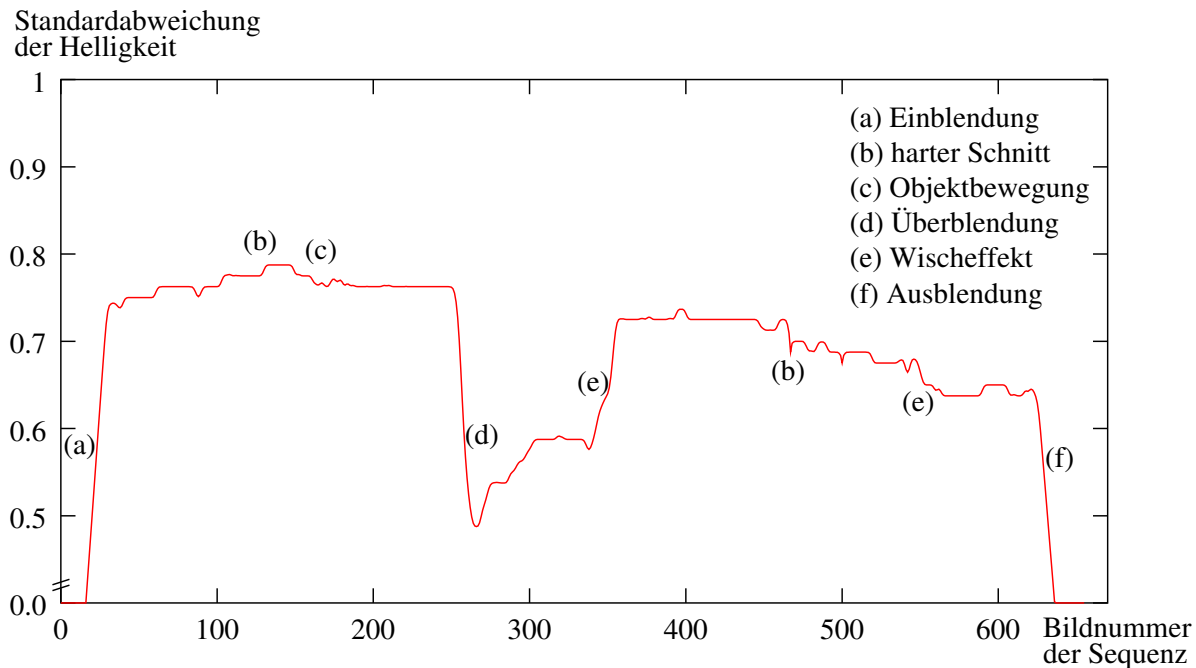


Abbildung 2.5: Die mit einem Gaußfilter geglättete Standardabweichung der Helligkeitswerte eines Bildes ermöglicht die Erkennung von Ein-, Aus- und Überblendungen. Harte Schnitte und Wischeffekte können nicht erkannt werden.

Die Summe beider Distanzmaße ermöglicht die Erkennung weicher Schnitte entsprechend den Gleichungen 2.8 und 2.9.

Fehlerhafte Klassifikationen treten insbesondere bei schnellen Kamera- oder Objektbewegungen auf. Der Bildinhalt ist in diesen Kameraeinstellungen häufig unscharf, so dass die Standardabweichung der Pixelwerte sinkt. Die Unschärfe entsteht während der Aufnahme und bei der Kompression des digitalen Videos. Starke Bewegungen verursachen deutliche Unterschiede in aufeinander folgenden Bildern. Um die Bitrate zu beschränken, werden die Blöcke im Bild stärker quantisiert, so dass insbesondere bei starken Bewegungen scharfe Kanten verloren gehen.

Ein wesentlicher Vorteil der Schnitterkennung durch Analyse der Standardabweichung liegt in der geringen Komplexität der Berechnung. Das Verfahren eignet sich insbesondere in Kombination mit anderen Verfahren zur Erkennung weicher Schnitte, da es schnell und zuverlässig eine Auswahl möglicher Ein-, Aus- und Überblendungen liefert. Für harte Schnitte ist der Ansatz nicht geeignet, da die grundlegende Annahme, dass sich die Standardabweichung zweier Bilder aus unterschiedlichen Kameraeinstellungen signifikant unterscheidet, häufig nicht zutrifft.

2.2.4 Kantenbasierte Verfahren zur Schnitterkennung

Ein wesentlicher Nachteil der Schnitterkennung mit Pixeldifferenzen oder Histogrammen sind die hohen Fehlerraten bei Helligkeitsschwankungen. Kantenbasierte Verfahren liefern insbesondere bei Helligkeitsschwankungen zuverlässigere Klassifikationsergebnisse [348]. Im Folgenden werden die *Kantenänderungsrate* und der *kantenbasierte Kontrast* näher betrachtet [317, 462, 579, 580].

Zur Berechnung der *Kantenänderungsrate* (engl. *edge change ratio* bzw. *edge change fraction*) werden die Kanten in zwei Bildern eines Videos mit Hilfe des *Canny*-Kantendetektors ermittelt [70, 71]. Der als *Hysterese* benannte Schritt des Algorithmus markiert starke Kanten und zusätzlich alle schwachen Kanten, die mit einer starken Kante verbunden sind. Zwei Schwellwerte definieren, ab wann eine Kante als schwache bzw. starke Kante zählt. Obwohl es möglich ist, aus einem Bild die beiden Schwellwerte zuverlässig zu schätzen, dürfen sich zur Berechnung der Kantenänderungsrate die verwendeten Schwellwerte innerhalb eines Videos nicht ändern. Insbesondere bei Ein- und Ausblendungen würden sonst auch in fast monochromen Bildern viele Kanten erkannt werden, die überwiegend Rauschen im Bild repräsentieren. Ein fester Schwellwert für ein Video liefert dagegen vergleichbare Kantenbilder.

Kantenpixel, die im ersten aber nicht im zweiten Bild enthalten sind, werden als *ausgehende Kantenpixel* bezeichnet, die im zweiten Bild neu hinzukommenden Kantenpixel als *eingehende Kantenpixel*. $E_{out}(i)$ und $E_{in}(j)$ speichern die Anzahl der aus- und eingehenden Kantenpixel der Bilder i und j . $\rho_{out}(i)$ und $\rho_{in}(j)$ spezifizieren den Anteil der ausgehenden und eingehenden Kantenpixel zur gesamten Anzahl der Kantenpixel S_i eines Bildes i . Die Kantenänderungsrate $ECR_{i,j}$ für die beiden Bilder i und j ist definiert als:

$$\rho_{out}(i) = \frac{E_{out}(i)}{S_i} \quad (2.15)$$

$$\rho_{in}(j) = \frac{E_{in}(j)}{S_j} \quad (2.16)$$

$$ECR_{i,j} = \max \{ \rho_{out}(i), \rho_{in}(j) \}. \quad (2.17)$$

Da die Kantenänderungsrate innerhalb einer Überblendung mit der zeitlichen Entfernung zweier Bilder zunimmt und sie zwischen zwei benachbarten Bildern ähnliche Werte besitzt, ist eine Erkennung von Ein-, Aus- und Überblendungen mit Hilfe der Gleichungen 2.8 und 2.9 möglich. Schon kleine Kamera- oder Objektbewegungen können die Anzahl der eingehenden und ausgehenden Kanten signifikant erhöhen. Zur Reduktion der Fehler wird die Kamerabewegung

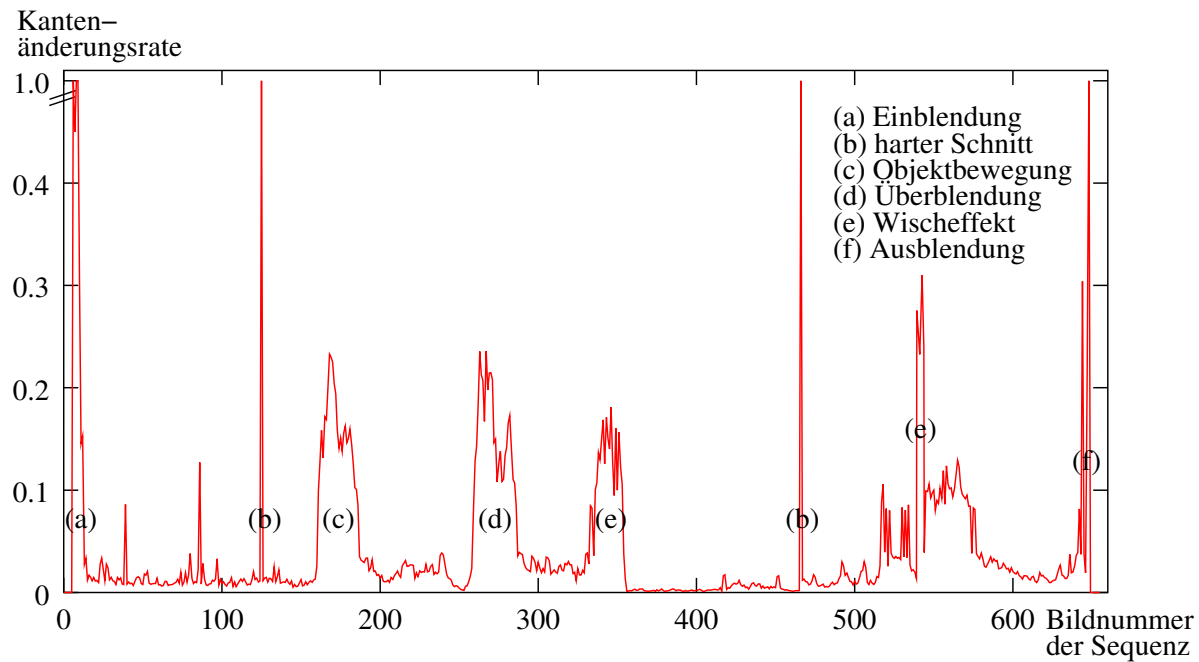


Abbildung 2.6: Änderung der Werte der Kantenänderungsrate in einer Videosequenz

ermittelt und kompensiert (vgl. Kapitel 3), so dass die Kanten des Hintergrundes beider Bilder an ähnlichen Positionen liegen. Trotz des Ausgleichs der Kamerabewegung können Kanten zweier Bilder geringfügig verschoben sein. Zur Reduktion der Fehler werden die Kanten eines Bildes mit Hilfe des Dilatationsoperators verbreitert [457, 467]. $E_{out}(i)$ zählt die Kantenpixel des Kantenbildes i , die nicht im dilatierten Kantenbild j vorkommen, $E_{in}(j)$ die des Kantenbildes j ohne die im dilatierten Kantenbild i auftretenden Kanten.

Ein einzelner hoher Wert der Kantenänderungsrate ist ein Indikator für einen harten Schnitt, wohingegen bei weichen Schnitten mehrere zusammenhängende leicht erhöhte Werte auftreten. Das Verhältnis der eingehenden zu den ausgehenden Kanten gibt einen Hinweis auf die Art des Schnittes: Während einer Ausblendung oder der ersten Hälfte einer Überblendung verschwinden Kanten, und die Werte für ρ_{out} sind größer als ρ_{in} , wogegen bei einer Einblendung und in der zweiten Hälfte einer Überblendung mehr eingehende als ausgehende Kanten auftreten ($\rho_{in} > \rho_{out}$). Abbildung 2.6 verdeutlicht die Änderung der Werte der Kantenänderungsrate innerhalb einer Videosequenz.

Ein wesentlicher Nachteil bei der Schnitterkennung mit Hilfe der Kantenänderungsrate ist die sehr hohe Anzahl an fehlerhaft erkannten Schnitten. Die meisten Fehlklassifikationen können auf Objektbewegungen zurückgeführt werden, da nur die Kamerabewegung ausreichend gut kompensiert wird. Kameraschwenks oder Zoomeffekte beeinflussen die Kantenänderungsrate

nur unwesentlich, und geringe Fehler bei der Schätzung der Parameter des Kameramodells haben durch die Dilatation keine signifikante Auswirkung.

Ein weiteres Maß zur Erkennung von Ein-, Aus- und Überblendungen ist der *kantenbasierte Kontrast* (engl. *Edge-based Contrast*) [315]. Dazu wird aus einem Kantenbild I ein aggregierter Wert für schwache Kanten w_I und starke Kanten s_I berechnet:

$$w_I = \sum_{x,y} \begin{cases} I(x,y) & \text{falls } \theta_w \leq I(x,y) < \theta_s, \\ 0 & \text{sonst} \end{cases} \quad (2.18)$$

$$s_I = \sum_{x,y} \begin{cases} I(x,y) & \text{falls } I(x,y) \geq \theta_s, \\ 0 & \text{sonst.} \end{cases} \quad (2.19)$$

Die Schwellwerte θ_w und θ_s legen fest, ab wann eine Kante als schwache oder starke Kante zählt. Der kantenbasierte Kontrast (EC) aggregiert die Werte für schwache und starke Kanten und ist definiert als:

$$EC = 1 + \frac{s_I - w_I - 1}{s_I + w_I + 1}, \quad EC \in [0, 2]. \quad (2.20)$$

Tabelle 2.2 verdeutlicht die Werteverteilung des kantenbasierten Kontrastes in Abhängigkeit von der Anzahl der starken und schwachen Kanten im Bild. Zu Beginn einer Ein- bzw. am Ende einer Ausblendung und in Kameraeinstellungen mit dunklen Bildinhalten ist die Anzahl der starken Kanten und damit der Wert für den kantenbasierten Kontrast sehr gering. Charakteristisch für eine Überblendung sind zunächst fallende Werte, die in der zweiten Hälfte der Überblendung wieder ansteigen. Der kantenbasierte Kontrast weist somit ähnliche Eigenschaften wie die Varianz der Helligkeitswerte im Bild auf. Als Distanzmaß wird die Summe der Gleichungen 2.13 und 2.14 verwendet.

Ein wesentlicher Vorteil der Erkennung weicher Schnitte mit dem kantenbasierten Kontrast ist der geringe Einfluss der Kamera- bzw. Objektbewegung und die geringe Komplexität der Berechnung. Nur bei schnellen Bewegungen treten höhere Fehlerraten auf, da das Bild häufig an Schärfe verliert. Es sinkt die Anzahl der starken Kanten im Bild, so dass verstärkt Bewegungen als Überblendungen klassifiziert werden. Insbesondere in Kombination mit anderen Verfahren liefert der kantenbasierte Kontrast schnell und zuverlässig eine Auswahl möglicher Ein-, Aus- und Überblendungen.

Anteil starker und schwacher Kanten	EC
$s_I = 0$	0
$s_I < w_I$	$0 < EC < 1$
$s_I \approx w_I > 0$	1
$s_I > w_I$	$1 < EC < 2$
$s_I \gg w_I$	2

Tabelle 2.2: Auswirkung der Anzahl schwacher und starker Kanten auf den kantenbasierten Kontrast

2.2.5 Verbesserung der Schnitterkennung durch Bewegungsanalyse

Die bisher vorgestellten Verfahren sind fehleranfällig bei starken Kamera- oder Objektbewegungen im Video, so dass viele Kameraeinstellungen mit starker Bewegung als Schnitt klassifiziert werden. Durch die Analyse der Kamerabewegung ist es möglich, die Anzahl der fehlerhaft erkannten Schnitte zu reduzieren [80, 130, 422].

Die durch die Kamerabewegung erzeugte Änderung der Position der Pixel im Bild kann mit Hilfe eines affinen oder perspektivischen Modells beschrieben werden (vgl. Kapitel 3). Aus den Parametern des Modells werden Beschreibungen für mögliche Kamerabewegungen wie Kameraschwenks oder Zoomoperationen abgeleitet. Eine kontinuierliche Kamerabewegung über mehrere Bilder deutet darauf hin, dass kein Schnitt innerhalb dieser Bilder vorhanden ist. Auch die Analyse der Bewegung von Objekten im Bildvordergrund (vgl. Kapitel 4) kann einen Hinweis auf fehlerhaft erkannte Schnitte liefern. Insbesondere bei einer kontinuierlichen Bewegung eines Objektes kann ein Schnitt ausgeschlossen werden.

Bewegungsbasierte Verfahren eignen sich insbesondere in Kombination mit anderen Verfahren zur Schnitterkennung. Erfolgreich werden diese Verfahren mit pixelbasierten Verfahren [394, 519], Histogrammen [197, 469, 584] und kantenbasierten Verfahren [57, 581] kombiniert und können die Ergebnisse der Schnitterkennung signifikant verbessern.

2.3 Experimentelle Ergebnisse

Im Rahmen der TRECVID-Konferenz [292] werden umfangreiche Sammlungen von Videos zur Verfügung gestellt, um Schnitterkennungsverfahren mit einer einheitlichen Datenbasis vergleichen zu können. Bei dem überwiegenden Teil der Daten handelt es sich um Dokumentationen und Nachrichtensendungen, die im Fernsehen nur einen relativ geringen Teil des Programms ausmachen. Aus dem Jahr 2005 umfasst das Videomaterial von TRECVID 169

Bezeichnung	Dauer [min]	Anzahl harter Schnitte	Anzahl Ein- und Aus- blendungen	Anzahl Über- blendungen	Anzahl Wisch- effekte
Dokumentation	12	86	9	1	0
Nachrichtensendung	15	109	0	9	2
Spielfilm	17	275	0	11	0
Talkshow	16	134	0	0	0
Serie	15	221	2	7	0
Zeichentrickfilm	10	175	3	20	1
Sportsendung	14	107	0	12	0
Musikclip	11	192	21	79	0
Werbung	11	305	8	29	2
Summe	121	1604	43	168	5

Tabelle 2.3: Verteilung der Schnitte in den ausgewählten Videosequenzen

Stunden Nachrichtensendungen, zu denen noch vier wissenschaftliche Videos hinzugenommen wurden [403].

Innerhalb der experimentellen Ergebnisse soll eine möglichst allgemeine Aussage über die Qualität der Schnitterkennungsverfahren getroffen werden. Ein wesentlicher Nachteil bei der Analyse der Schnitterkennung mit den TRECVID-Videos besteht darin, dass fast ausschließlich Nachrichtensendungen vorhanden sind und Aussagen über die Schnitterkennungsverfahren für andere Genres nur bedingt möglich sind.

Um eine allgemeinere Aussage über die Qualität eines Verfahrens zur Erkennung von Schnitten in Videos zu ermöglichen, haben wir neun Videosequenzen aus dem Fernsehen¹ mit einer Länge zwischen 10 und 17 Minuten aus unterschiedlichen Genres zusammengestellt und zunächst theoretische Obergrenzen für ausgewählte Schnitterkennungsverfahren analysiert. Die Längen der einzelnen Videos und die Anzahl der Schnitte sind in Tabelle 2.3 aufgelistet. Die Schwellwerte werden in einem zweiten Analyseschritt verwendet, um tatsächliche Erkennungsraten für unbekannte Videos zu ermitteln. Eine allgemein gültige Aussage über die Qualität der Erkennung von Wischeffekten ist aufgrund ihrer geringen Anzahl in den untersuchten Videosequenzen nicht möglich.

Die beiden Maße *Präzision* P (engl.: *precision*) und *Vollständigkeit eines Suchergebnisses* V (engl.: *recall*) liefern Werte für die Qualität eines Verfahrens zur Erkennung von Schnitten. Sie sind definiert als:

¹MPEG-2 Videos in PAL-Auflösung, Bildwiederholrate: 25 Bilder/s, Bitrate: 4,5 MBit/s

$$P = \frac{C}{C+F} \in [0, 1] \quad (2.21)$$

$$V = \frac{C}{C+M} \in [0, 1]. \quad (2.22)$$

C und F bezeichnen die Anzahl der korrekt bzw. fehlerhaft erkannten Schnitte. Deren Summe $(C + F)$ entspricht der gesamten Anzahl Schnitte, die der Algorithmus ermittelt hat. M zählt die Schnitte, die nicht erkannt werden konnten, so dass $(C + M)$ der tatsächlichen Anzahl der Schnitte des Videos entspricht. Der maximale Wert von eins für die Präzision bedeutet, dass es sich bei allen erkannten Schnitten um echte Schnitte des Videos handelt. Sind in der Menge der erkannten Schnitte alle Schnitte des Videos enthalten, so erreicht die Vollständigkeit den maximalen Wert von eins.

Wird nur eines der beiden Maße betrachtet, so ist keine Aussage über die Qualität eines Verfahrens für die Schnitterkennung möglich. Da die Vollständigkeit den maximalen Wert erreicht, wenn kein Schnitt ausgelassen wird, könnte zur Maximierung der Vollständigkeit zwischen jedem Bild ein harter Schnitt gewählt werden. Andererseits kann jedes Verfahren so angepasst werden, dass nur die sehr eindeutigen Schnitte als solche klassifiziert werden. Der Extremfall wäre die Auswahl eines einzelnen Schnittes in einem Video, so dass mit hoher Wahrscheinlichkeit die Präzision den maximalen Wert erreicht. Da eine getrennte Optimierung keine sinnvollen Ergebnisse liefert, hat sich das $F1$ –Maß [44] als Kombination von Präzision und Vollständigkeit zur Beurteilung der Qualität von Schnitterkennungsverfahren durchgesetzt:

$$F1 = 2 \cdot \frac{P * V}{P + V} \in [0, 1] \quad \text{für } P, V \neq 0. \quad (2.23)$$

2.3.1 Theoretische Obergrenzen für die Erkennung harter Schnitte

Für den Vergleich der Schnitterkennungsverfahren haben wir die in Tabelle 2.4 aufgeführten Verfahren implementiert und für die analysierten Videosequenzen zunächst theoretische Obergrenzen für die Qualität der unterschiedlichen Schnitterkennungsverfahren ermittelt. Jedes Verfahren liefert den Differenzwert $D_{i,i+1}$ für zwei benachbarte Bilder. Falls die Werte den Schwellwert $T_{i,i+1}$ übersteigen, wird zwischen den Bildern i und $i + 1$ ein harter Schnitt erkannt. Ein optimaler Schwellwert wurde im Vorfeld für jedes einzelne Verfahren manuell bestimmt, so dass für die ausgewählten Videos die angegebenen Ergebnisse als theoretisches Optimum für den $F1$ -Wert angesehen werden können. Da der jeweils optimale Schwellwert nicht

Verfahren	Präzision	Vollständigkeit	F1	Rechenzeit
Summe absoluter Differenzen	85,2 %	82,7 %	83,9 %	0,86
Kantenänderungsrate	76,1 %	86,5 %	81,0 %	7,78
Histogramm	60,4 %	79,2 %	68,5 %	0,67
Durchschnittlicher Farbwert	56,9 %	68,2 %	62,0 %	0,67
Kontrast	55,7 %	68,9 %	61,6 %	0,76
Bewegungsvektoren	25,6 %	92,4 %	40,0 %	2,81

Tabelle 2.4: Theoretische Obergrenzen für die Erkennung harter Schnitte in den analysierten Videos. Die Rechenzeit jedes Verfahrens ist als Faktor im Vergleich zur Länge des Videos angegeben.

automatisch ermittelt werden kann, sind im realen Einsatz Abweichungen von den optimalen Ergebnissen zu erwarten. Dies wird auch bei der Analyse der unbekannten Videosequenzen in Abschnitt 2.3.5 deutlich.

In Tabelle 2.4 sind die Präzision, Vollständigkeit und der $F1$ –Wert für die unterschiedlichen Schnitterkennungsverfahren angegeben. Besonders gut eignet sich die Summe der absoluten Differenzen und die Kantenänderungsrate zur Erkennung harter Schnitte. Der Aufwand zur Berechnung der Kantenänderungsrate liegt dabei fast um den Faktor zehn höher. Verfahren, welche die Bildinformationen auf einen einzelnen Wert aggregieren, führen zu deutlich ungenaueren Klassifikationsergebnissen. Beispiele hierfür sind der durchschnittliche Farbwert und der Kontrast. Histogrammbasierte Verfahren erreichen gute Klassifikationsergebnisse, wobei hohe Fehlerraten bei Helligkeitsschwankungen und starken Bewegungen auftreten.

Die Analyse der Änderung der durchschnittlichen Länge der Bewegungsvektoren ermöglicht trotz eines guten Wertes für die Vollständigkeit keine zuverlässige Erkennung harter Schnitte, da in Kameraeinstellungen mit starken Objektbewegungen viele Schnitte fehlerhaft klassifiziert werden und so die Präzision einen sehr geringen Wert annimmt. Im folgenden Abschnitt wird jedoch deutlich, dass durch geschickte Kombination zweier Verfahren die Analyse der Bewegungen die Schnitterkennungsergebnisse deutlich verbessert.

2.3.2 Optimierungen zur Erkennung harter Schnitte

Zwei Verfahren zur Verbesserung der Klassifikationsergebnisse werden im Folgenden vorgeschlagen. Das erste Verfahren verwendet eine ähnliche Idee wie Yeo et al. [569], die adaptive Schwellwerte zur Erkennung harter Schnitte genutzt haben, um die signifikant höheren Fehlerraten in Kameraeinstellungen mit starken Bewegungen zu reduzieren. Zur Erkennung eines harten Schnittes wird für jedes Bild i ein durchschnittlicher Differenzwert $D_{i,i+1}^{avg}$ berechnet, in

Verfahren	Präzision		Vollständigkeit		F1	
	(a)	(b)	(a)	(b)	(a)	(b)
Summe absoluter Differenzen	94,4 %	94,7 %	94,2 %	94,7 %	94,3 %	94,7 %
Kantenänderungsrate	82,8 %	89,8 %	92,2 %	97,2 %	87,2 %	93,3 %
Histogramm	81,4 %	84,6 %	89,0 %	89,5 %	85,0 %	87,0 %
Durchschnittlicher Farbwert	74,1 %	76,0 %	76,6 %	76,6 %	75,3 %	76,3 %
Kontrast	72,7 %	73,4 %	74,6 %	77,5 %	73,6 %	75,4 %
Bewegungsvektoren	49,8 %	—	73,0 %	—	59,2 %	—

Tabelle 2.5: Theoretische Obergrenzen für die Erkennung harter Schnitte in den analysierten Videosequenzen unter Berücksichtigung adaptiver Schwellwerte (a) und in Kombination mit Bewegungsvektoren (b)

den die Differenzwerte der benachbarten Bilder einfließen:

$$D_{i,i+1}^{avg} = \frac{1}{N} \sum_{j=i-\frac{N}{2}, j \neq i}^{i+\frac{N}{2}} D_{j,j+1}. \quad (2.24)$$

N spezifiziert die Anzahl der benachbarten Bilder, aus denen der Durchschnittswert berechnet wird. Der adaptive Schwellwert wird definiert durch $T_{i,i+1} = D_{i,i+1}^{avg} + T$. Übersteigt die Distanz zwischen den Bildern i und $i + 1$ die Summe aus dem global festgelegten Schwellwert T und $D_{i,i+1}^{avg}$, so wird ein harter Schnitt klassifiziert.

In der Tabelle 2.5 (a) sind die Klassifikationsergebnisse unter Berücksichtigung der durchschnittlichen Differenzwerte der benachbarten Bilder angegeben. Das Verfahren verbessert den $F1$ –Wert für alle Merkmale, wobei die Rechenzeit nur minimal zunimmt (weniger als 0,1 Prozent). Eine deutliche Steigerung der $F1$ –Werte kann bei Histogrammen und der Summe der absoluten Differenzen beobachtet werden. Die Summe der absoluten Differenzen erreicht einen $F1$ –Wert von über 94 Prozent, wobei die Ergebnisse der Kantenänderungsrate und Histogramme mit deutlichem Abstand folgen; sie liegen bei 87 bzw. 85 Prozent.

Die zweite Verbesserungsmöglichkeit der analysierten Algorithmen erfolgt durch eine geschickte Kombination zweier Schnitterkennungsverfahren, wobei beim ersten Verfahren die Parameter so spezifiziert werden sollten, dass die Vollständigkeit einen hohen Wert annimmt, um anschließend den $F1$ –Wert mit dem zweiten Verfahren zu maximieren. So können in einem ersten Schritt alle Bereiche des Videos erkannt und verworfen werden, in denen mit hoher Wahrscheinlichkeit keine Schnitte enthalten sind. Anschließend wird für die restlichen Bilder die Schnitterkennung mit einem zweiten Verfahren durchgeführt. Obwohl das auf Bewegungs-

Bezeichnung	SAD	ECR	HD	Farbe	Kontrast
Dokumentation	0.06	0.04	0.35	0.020	0.019
Nachrichtensendung	0.03	0.10	0.20	0.003	0.004
Spielfilm	0.08	0.46	0.08	0.003	0.003
Talkshow	0.05	0.48	0.09	0.009	0.002
Serie	0.07	0.34	0.25	0.007	0.002
Zeichentrickfilm	0.08	0.24	0.38	0.008	0.011
Sportsendung	0.03	0.21	0.15	0.008	0.008
Musikclip	0.05	0.25	0.22	0.020	0.020
Werbung	0.09	0.33	0.24	0.025	0.017
Durchschnitt	0,06	0.27	0.22	0.011	0.010

Tabelle 2.6: *Optimale Schwellwerte für die Erkennung harter Schnitte in den analysierten Testvideos: Summe der absoluten Differenzen (SAD), Kantenänderungsrate (ECR), Histogrammdifferenz (HD), durchschnittliche Farbe und Kontrast.*

vektoren basierende Verfahren bei der Klassifikation von harten Schnitten schlechte Ergebnisse liefert, ist es mit diesem Verfahren möglich, viele Bereiche des Videos zu identifizieren, in denen keine harten Schnitte enthalten sind. So kann bei einer geringen Länge der Bewegungsvektoren angenommen werden, dass kein harter Schnitt zwischen zwei Bildern liegt, da sonst zufällig verteilte und somit auch längere Bewegungsvektoren auftreten würden. Trotz längerer Bewegungsvektoren wird bei einer kontinuierlichen Kamerabewegung, wie sie beispielsweise bei einem Kameraschwenk auftritt, ebenfalls ein harter Schnitt ausgeschlossen.

Die Analyse der Kamerabewegung liefert eine Auswahl möglicher harter Schnitte. Eine Maximierung der Vollständigkeit verhindert, dass viele echte Schnitte aussortiert werden. Für die getesteten Videosequenzen wurde der Schwellwert so festgelegt, dass die Vollständigkeit maximal ist und in der Menge der ausgewählten Bilder alle Schnitte enthalten sind. Die Präzision bei der Verwendung der Bewegungsvektoren sinkt dadurch auf einen Wert von unter 14 Prozent und entspricht dem Anteil der echten Schnitte innerhalb der ausgewählten Bilder. Die tatsächliche Schnitterkennung erfolgt anschließend mit einem zweiten Verfahren.

Obwohl Bewegungsvektoren als alleiniges Merkmal zur Schnitterkennung nicht geeignet sind, können sie in Kombination mit einem weiteren Verfahren die Klassifikationsergebnisse wesentlich verbessern. Aus den Ergebnissen in Tabelle 2.5 (b) wird deutlich, dass jede Kombination zu einer Verbesserung der $F1$ –Werte führt. Insbesondere bei einer Kombination der Bewegungsvektoren mit der Kantenänderungsrate kann der $F1$ –Wert um mehr als sechs Prozent gesteigert werden.

Für die Berechnung der Klassifikationsergebnisse aus Tabelle 2.5 wurden ebenfalls optima-

le Schwellwerte für die einzelnen Testvideos bestimmt, die in Tabelle 2.6 angegeben sind. Anhand der Werte wird deutlich, dass trotz der sehr unterschiedlichen Arten von Videos die Schwellwerte nur geringfügig voneinander abweichen. Wie in Abschnitt 2.3.5 deutlich wird, liefert die Tabelle 2.6 gute Schätzwerte für die Schwellwerte der Schnitterkennungsverfahren. Falls das Genre unbekannt ist, liefert der durchschnittliche Wert eine Schätzung für die Schwellwerte. Neben den Schwellwerten sind für einzelne Verfahren noch weitere Parameter zu berücksichtigen: Der adaptive Schwellwert $D_{i,i+1}^{avg}$ berücksichtigt als Differenzwerte bei Histogrammdifferenzen $N = 4$, bei der Summe der absoluten Differenz und bei Farbdifferenzen $N = 6$ sowie bei der Kantenänderungsrate und dem Kontrast $N = 8$ benachbarte Bilder. Die Berechnung der Histogrammdifferenzen erfolgt mit der L_1 -Norm anhand von YUV -Bildern, wobei vier Bits zur Beschreibung der Helligkeit und jeweils drei Bits für die Farbwerte verwendet werden.

2.3.3 Theoretische Obergrenzen für die Erkennung weicher Schnitte

Im Vergleich zu harten Schnitten treten bei der Erkennung und korrekten Klassifikation eines weichen Schnittes wesentlich höhere Fehlerraten auf. Die Änderungen zwischen zwei benachbarten Bildern sind innerhalb eines weichen Schnittes sehr gering, so dass die Merkmale über einen längeren Zeitraum analysiert werden müssen. Eine Unterscheidung zwischen Objekt- oder Kamerabewegung und einem weichen Schnitt ist aufgrund der Ähnlichkeit der Differenzwerte nicht immer möglich. Eine weitere Schwierigkeit liegt in der exakten Erkennung des Start- und Endpunktes eines weichen Schnittes, da auch ein Mensch die genaue Position nicht immer eindeutig bestimmen kann. Ein weicher Schnitt gilt für die folgenden Ergebnisse als korrekt erkannt, wenn mehr als die Hälfte der Bilder eines weichen Schnittes übereinstimmen.

Die Qualität der Erkennung von Ein-, Aus- und Überblendungen wird für die ausgewählten Videosequenzen analysiert. Ein- und Ausblendungen können als Spezialfall einer Überblendung angesehen werden, bei der die erste bzw. zweite Kameraeinstellung nur monochrome Bilder enthält, so dass alle Verfahren zur Erkennung von Überblendungen auch Ein- oder Ausblendungen erkennen. Da der Anteil der Wischeffekte in den analysierten Videos sehr gering und somit eine repräsentative Aussage über die Erkennungsqualität nicht möglich ist, werden diese nicht weiter betrachtet.

Der Kontrast, Histogrammdifferenzen, der kantenbasierte Kontrast und die Kantenänderungsrate werden hinsichtlich ihrer Eignung zur Erkennung einer Überblendung analysiert. Es ist

nicht möglich, einen einzelnen Differenzwert mit einem Schwellwert zu vergleichen, um eine Überblendung zu erkennen. Vielmehr werden die Differenzwerte über einen Zeitraum von mehreren Bildern entsprechend den Gleichungen 2.3, 2.8 und 2.9 analysiert. Damit ein weicher Schnitt vorliegt, müssen die Distanzen zwischen dem ersten und letzten Bild einer Überblendung sehr groß sein, die Distanzen benachbarter Bilder ähnliche Werte annehmen und die Distanzen mit zunehmendem zeitlichen Abstand zweier Bilder ansteigen.

2.3.4 Optimierungen zur Erkennung weicher Schnitte

Zwei Verbesserungen werden im Folgenden vorgeschlagen, um mit Hilfe der Kantenänderungsrate und Histogrammdifferenzen bessere Klassifikationsergebnisse zu erzielen. Um Überblendungen mit Hilfe der Kantenänderungsrate zu erkennen, wurden in früheren Ansätzen erhöhte Werte innerhalb aufeinander folgender Bilder gesucht [317, 580]. Obwohl diese Vorgehensweise einen großen Teil der Überblendungen erkennt und einen Wert für die Vollständigkeit von 65 Prozent erreicht, treten insbesondere bei Objektbewegungen viele fehlerhafte Klassifikationen und somit ein geringer Wert für die Präzision auf. Der im Folgenden vorgestellte modifizierte Wert für die Kantenänderungsrate verbessert die Klassifikationsergebnisse deutlich, indem durch harte Schnitte verursachte Fehler ausgefiltert werden.

Die Differenzwerte der Kantenänderungsrate sind während einer Überblendung leicht erhöht, und bei einem harten Schnitt tritt ein einzelner stark ausgeprägter Wert auf. Die Summe der Werte der Kantenänderungsraten innerhalb einer Umgebung von N Bildern abzüglich des maximalen Wertes innerhalb dieser Umgebung liefert den *modifizierten Wert für die Kantenänderungsrate* M_i^{ECR} , der bei harten Schnitten niedrige und bei weichen Schnitten hohe Werte annimmt:

$$M_i^{ECR} = \sum_{j=i-\frac{N}{2}}^{i+\frac{N}{2}} ECR_{j,j+1} - \max \left\{ ECR_{j,j+1} : j = i - \frac{N}{2} \dots i + \frac{N}{2} \right\}. \quad (2.25)$$

Übersteigt die Kantenänderungsrate M_i^{ECR} einen Schwellwert, so wird eine Überblendung klassifiziert. Einen weiteren Hinweis liefert das Verhältnis der eingehenden zu den ausgehenden Kanten: In der ersten Hälfte einer Überblendung liegt die Zahl der ausgehenden Kanten über der Anzahl der eingehenden Kanten, in der zweiten Hälfte ist das Verhältnis umgekehrt. Das zweite neue Distanzmaß verwendet Histogrammdifferenzen zur Erkennung von Überblendungen. Innerhalb eines weichen Schnittes sind die Histogrammdifferenzen von benachbarten

Verfahren	Präzision		Vollständigkeit		F1	
	(a)	(b)	(a)	(b)	(a)	(b)
Kantenänderungsrate	45,0 %	75,8 %	43,1 %	38,9 %	44,0 %	51,4 %
Histogramm	58,3 %	66,7 %	52,1 %	70,1 %	55,0 %	68,4 %
Kontrast	54,2 %	60,4 %	59,2 %	66,8 %	56,6 %	63,5 %
Kantenbasierter Kontrast	46,1 %	55,2 %	37,9 %	46,4 %	41,6 %	50,4 %
Kontrast	97,7 %		74,4 %		84,5 %	
Kantenbasierter Kontrast	93,0 %		72,1 %		81,2 %	

Tabelle 2.7: *Oben: Klassifikationsergebnisse für Überblendungen (a) und Verbesserung der Ergebnisse durch Entfernung automatisch erkannter harter Schnitte (b). Unten: Klassifikationsergebnisse für Aus- und Einblendungen*

Bildern sehr gering, so dass eine zuverlässige Erkennung von Überblendungen nicht direkt möglich ist. Die im Folgenden vorgeschlagene modifizierte Histogrammdifferenz liefert ein geeignetes Maß zur Erkennung von Überblendungen. Wird die Histogrammdifferenz nicht zwischen benachbarten Bildern, sondern zwischen jedem n -ten Bild des Videos berechnet, so treten hohe Differenzwerte bei weichen Schnitten auf. Die Klassifikationsergebnisse sind jedoch nicht sehr zuverlässig, da auch harte Schnitte und längere Kamerabewegungen zu erhöhten Werten führen. Der Einfluss harter Schnitte kann durch die Verwendung des modifizierten Histogrammdifferenzwertes M_i^{HD} vermieden werden:

$$M_i^{HD} = HD_{i-\frac{N}{2}, i+\frac{N}{2}} - \max \left\{ HD_{j,j+1} : j = i - \frac{N}{2} \dots i + \frac{N}{2} \right\}. \quad (2.26)$$

$HD_{j,j+1}$ bezeichnet die Histogrammdifferenz zweier benachbarter Bilder j und $j + 1$, die im Fall eines harten Schnittes einen großen Wert annimmt. Hohe Histogrammdifferenzwerte $HD_{i-\frac{N}{2}, i+\frac{N}{2}}$ zwischen Bild $i - \frac{N}{2}$ und $i + \frac{N}{2}$ treten bei harten und weichen Schnitten auf. Die *modifizierte Histogrammdifferenz* M_i^{HD} enthält nur innerhalb eines weichen Schnittes einen hohen Differenzwert.

Für die Klassifikationsergebnisse in Tabelle 2.7 (a) wurden optimale Schwellwerte zur Maximierung der $F1$ -Werte der analysierten Videosequenzen verwendet. Sowohl der Kontrast als auch die Histogrammdifferenz liefern gute Ergebnisse, die jedoch nicht an die Klassifikationsergebnisse für harte Schnitte heranreichen. Eine zusätzliche Verbesserung der Ergebnisse ist möglich, indem zunächst die automatisch erkannten harten Schnitte identifiziert und entfernt werden (Tabelle 2.7 (b)). Ein großer Teil der durch harte Schnitte verursachten Fehler kann so vermieden werden, wobei der gute $F1$ -Wert von über 94 Prozent bei der Erkennung von har-

Verfahren	Optimale Parameter
Kantenänderungsrate	$N=8, M_i^{ECR} > 5,5$
Modifizierte Histogramme	$N=12, M_i^{HD} = 1,59$
Kontrast	$N=6$, Größe der Maske für die Gaußglättung: 5, $T_{i,j} = 3,8$
Kantenbasierter Kontrast	$N=6, T_{i,j} = 7,2, \theta_w = 50, \theta_w = 100$

Tabelle 2.8: *Optimale Parameter für weiche Schnitte der analysierten Testvideos*

ten Schnitten garantiert, dass nur wenige weiche Schnitte entfernt werden. Die im Vergleich zur Erkennung von harten Schnitten immer noch hohen Fehlerraten werden durch Objekt- und Kamerabewegungen verursacht, da keine zuverlässige Unterscheidung zwischen Bildänderungen, die durch Bewegungen oder Überblendungen verursacht werden, möglich ist.

Um eine Ein- oder Ausblendung innerhalb der Menge der Überblendungen zu identifizieren, reicht es aus, den Kontrast bzw. den kantenbasierten Kontrast zu betrachten. Sinkt der Wert kontinuierlich unter eine bestimmte Grenze, so wird eine Ausblendung erkannt. Im unteren Bereich der Tabelle 2.7 wird deutlich, dass die Erkennung von Ein- und Ausblendungen sehr zuverlässig möglich ist und viele der Ein- und Ausblendungen korrekt identifiziert werden. Tabelle 2.8 gibt die optimalen Schwellwerte für die analysierten Videos bei der Erkennung weicher Schnitte an.

Zusammenfassend lässt sich festhalten, dass bei optimal gewählten Schwellwerten die Standardverfahren $F1$ –Werte von 83 bzw. 56 Prozent für harte und weiche Schnitte erreichen. Durch die Verbesserungen der Verfahren ist eine Steigerung bei harten Schnitten auf über 94 Prozent möglich. Für die ausgewählten Videosequenzen erreicht der $F1$ –Wert bei Überblendungen beim besten Verfahren 68 Prozent.

2.3.5 Klassifikationsergebnisse für harte und weiche Schnitte

In einem zweiten Analyseschritt wird für eine weitere Zusammenstellung von Videosequenzen² die Qualität der Schnitterkennung ohne Kenntnis der für diese Sequenzen optimalen Schwellwerte analysiert. Es wurden neun Videos aus unterschiedlichen Genres und einer Länge von jeweils fünf Minuten aus dem Fernsehen aufgezeichnet. Insgesamt enthalten die Sequenzen 791 harte Schnitte, 74 Überblendungen sowie 11 Ein- bzw. Ausblendungen. Die Erkennungsraten für harte und weiche Schnitte werden anhand der aus den anderen Sequenzen ermittelten durchschnittlichen Schwellwerte entsprechend den Tabellen 2.6 und 2.8 bestimmt.

²DIVX–Videos, Bildauflösung: 352 x 288 Pixel, Bildwiederholrate: 25 Bilder/s, Bitrate: 1 MBit/s

Verfahren	Präzision	Vollständigkeit	F1
Summe absoluter Differenzen	95,7 %	96,0 %	95,8 %
Kantenänderungsrate	85,2 %	92,2 %	88,6 %
Histogramm	81,5 %	84,3 %	82,9 %
Durchschnittlicher Farbwert	81,3 %	80,8 %	81,0 %
Kontrast	80,2 %	75,3 %	77,7 %
Kantenänderungsrate	47,5 %	37,8 %	42,1 %
Histogramm	47,8 %	43,2 %	45,4 %
Kontrast	42,7 %	43,2 %	43,0 %
Kantenbasierter Kontrast	49,3 %	44,6 %	46,8 %
Kontrast	66,7 %	90,9 %	76,9 %
Kantenbasierter Kontrast	75,0 %	81,8 %	78,3 %

Tabelle 2.9: Klassifikationsergebnisse für harte Schnitte (oben), Überblendungen (Mitte) und Ein- bzw. Ausblendungen (unten)

Die Tabelle 2.9 gibt die Klassifikationsergebnisse für harte und weiche Schnitte an. Bei der Erkennung harter Schnitte sind bei einzelnen Verfahren zum Teil deutliche Abweichungen bei den Erkennungsraten zu beobachten. So sinken die *F1*-Werte bei der Kantenänderungsrate und bei Histogrammen um fünf Prozent. Dagegen führt die Analyse des Farbwertes zu deutlich besseren Ergebnissen. Auch bei der Verwendung der absoluten Pixeldifferenzen ist eine Steigerung des *F1*-Wertes von 94,7 auf 95,8 Prozent möglich.

Stärkere Abweichungen von den vorherigen Klassifikationsergebnissen treten bei der Erkennung weicher Schnitte auf. Deutlich geringere *F1*-Werte werden insbesondere bei Histogrammen und dem Kontrast erreicht. Auch der Wert für die Kantenänderungsrate fällt auf 42,1 Prozent. Lediglich die Ergebnisse beim kantenbasierten Kontrast erweist sich als sehr stabil. Bei Ein- oder Ausblendungen ist weiterhin eine sehr zuverlässige Schnitterkennung möglich, und die *F1*-Werte beider Verfahren erreichen fast 80 Prozent.

Obwohl sich die beiden Zusammenstellungen von Videos deutlich bezüglich der Aufnahmezeit, den Sendern, der Videokompression und der Bildauflösung unterscheiden, weichen die Klassifikationsergebnisse beider Gruppen nicht allzu deutlich voneinander ab. Die Ergebnisse für harte Schnitte sowie Ein- und Ausblendungen sind sehr ähnlich. Lediglich bei Überblendungen treten deutliche Unterschiede beim *F1*-Wert auf.

2.3.6 Schnitterkennung in historischen Videos

Im Rahmen des Projektes *European Chronicles Online* [451, 452] wurde ein komplexes Softwaresystem entwickelt, um Archive mit umfangreichen Sammlungen historischer Videos zu verwalten und sowohl den Archivaren als auch der Öffentlichkeit leichter zugänglich zu machen. Vom kulturellen Standpunkt aus betrachtet handelt es sich bei den im Rahmen des Projektes analysierten Videos um sehr wertvolle Filme, die das Leben und besondere Ereignisse aus den unterschiedlichen europäischen Ländern von Anfang des letzten Jahrhunderts bis heute dokumentieren.

Vier große Filmarchive³, die mehr als 100.000 Stunden historischer Filme aufbewahren, haben einen Teil ihrer Filme für das Projekt zur Verfügung gestellt. 4500 Videos aus den Jahren 1920 bis 1965 und mit einer Gesamtlänge von mehr als 200 Stunden wurden ausgewählt und im System gespeichert. Viele Algorithmen zur automatischen Analyse von Videos sind in das *European-Chronicles-Online*-System integriert und unterstützen die Archivare und Anwender beim Zugriff auf die Videos.

Anwendungen zur automatischen *Erzeugung einer Zusammenfassung eines Videos*, zur *Schnitterkennung*, zur *Objekterkennung* und zur *Gesichtserkennung* wurden im Rahmen dieser Arbeit entwickelt und sind Bestandteil des *European-Chronicles-Online*-Systems. Nach dem Einfügen eines neuen Videos werden die Algorithmen zur Analyse des Videos automatisch gestartet. Die Ergebnisse der Berechnungen werden als Metadaten oder im Fall einer Zusammenfassung als kurzes Video im *European-Chronicles-Online*-Archiv gespeichert und liefern den Anwendern zusätzliche Informationen über die Videos.

Die Bildqualität der historischen Schwarz-Weiß-Filme ist mit der Qualität aktueller Filme nicht vergleichbar, da die Lagerung der Filmrollen über mehrere Jahrzehnte und der mechanische Abrieb beim Abspielen der Filme mit den alten Projektoren zu vielen Bildfehlern geführt haben. Die wesentlichen Eigenschaften der historischen Videos können wie folgt charakterisiert werden:

- Bei den analysierten Videos handelt es sich um *Schwarz-Weiß-Filme*, so dass farbbasierte Merkmale für die Videoanalyse nicht geeignet sind. Für die Schnitterkennung ist eine Analyse des durchschnittlichen Farbwertes nicht möglich.
- Durch die Lagerung der Filmrollen und die geringe Qualität der historischen Kameras ist viel *Rauschen* in den Bildern enthalten, so dass bei der Schnitterkennung der Vergleich

³Instituto Luce (Italien), Memoriav (Schweiz), Netherlands Institute for Sound and Vision (Niederlande) und Institut Nationale de l'Audiovisuel (Frankreich)

Verfahren	F1-Wert für Farbvideos	F1-Wert für historische Videos
Summe absoluter Differenzen	94,7 %	86,0 %
Kantenänderungsrate	93,3 %	42,7 %
Histogramm (L_1 -Norm)	87,0 %	69,3 %
Durchschnittlicher Farb-/Helligkeitswert	76,3 %	65,8 %
Kontrast	75,4 %	64,8 %
Histogramm (<i>Earth-Movers</i> -Distanz)	87,1 %	77,4 %
Kombination der Verfahren	94,9 %	91,5 %

Tabelle 2.10: Klassifikationsergebnisse für harte Schnitte in historischen Videos

benachbarter Bilder zu einer deutlich höheren Fehlerrate führt. Als Folge liefern insbesondere kantenbasierte Verfahren bei den historischen Videos nur sehr unzuverlässige Ergebnisse.

- Durch Ermüdung des Filmmaterials und die veraltete Technik der Projektoren treten deutliche *Helligkeitsschwankungen* auf. Falls die für Farbvideos ermittelten Schwellwerte zur Analyse historischer Videos übernommen werden, sind nur sehr schlechte Klassifikationsergebnisse möglich. Es sollte insbesondere ein höherer Wert für N gewählt werden, der den adaptierten Schwellwert beeinflusst.
- Zur Digitalisierung werden die alten Filmrollen auf den historischen Projektoren abgespielt und mit modernen Kameras aufgezeichnet. Durch den mechanischen Filmtransport in den Projektoren sind viele historische Videos *verwackelt*, so dass die Fehlerrate bei der Analyse der Bewegungen deutlich steigt.
- Die Oberfläche der Filme ist durch mechanischen Abrieb und die Lagerung der Filmrollen teilweise stark beschädigt. Viele Videos enthalten *Bildfehler* in Form von *Streifen*, *Kratzern* sowie hellen oder dunklen *Flecken*, die sich deutlich vom Bildinhalt abheben. Durch die lokalen Bildfehler sinkt die Qualität der kantenbasierten Verfahren deutlich.

Zur Analyse der Qualität der Schnitterkennungsverfahren werden 40 historische Videos mit einer gesamten Länge von 196 Minuten ausgewählt. Die Videos enthalten fast ausschließlich harte Schnitte, da das manuelle Erzeugen eines weichen Schnittes mit großem Aufwand verbunden ist. Die ausgewählten Videos enthalten insgesamt 2544 harte Schnitte. Die durchschnittliche Länge der Kameraeinstellungen in den historischen Videodokumentationen ist mit weniger als fünf Sekunden sehr kurz.

Die geringe Bildqualität der historischen Videos hat großen Einfluss auf die Klassifikationsergebnisse und führt zu einer deutlichen Verringerung der Präzision und der Vollständigkeit bei der Schnitterkennung. In Tabelle 2.10 werden die Ergebnisse der Schnitterkennungsalgorithmen für Farbvideos und historische Videos verglichen. Im Vorfeld wurden anhand zufällig ausgewählter historischer Videos mit einer Länge von insgesamt 60 Minuten geeignete Schwellwerte ermittelt und diese für die eigentliche Erkennung verwendet. Bei der Erkennung harter Schnitte sinkt der $F1$ -Wert im Vergleich zu Farbvideos um 9 bis 51 Prozentpunkte. Die größte Abweichung von 93,3 auf 42,7 Prozent ist bei der Kantenänderungsrate zu beobachten, da durch Rauschen und Bildfehler ständig neue Kanten im Bild erscheinen bzw. bestehende Kanten verschwinden.

Zwei neue Verfahren, welche die Schnitterkennung speziell für historische Videos verbessern, werden im Folgenden vorgestellt. Bei dem ersten Verfahren wird statt der L_1 - oder L_2 -Norm ein für historische Videos besser geeignetes Distanzmaß zur Berechnung der Histogrammdifferenzen verwendet. Der zweite Ansatz kombiniert mehrere Schnitterkennungsverfahren speziell für historische Videos, da kantenbasierte Verfahren robust gegenüber Helligkeitsschwankungen sind und Histogramme auch bei starkem Rauschen und Bildfehlern gute Ergebnisse liefern.

Bei der Schnitterkennung mit Histogrammen können schon geringe Helligkeitsschwankungen zu deutlich höheren Fehlerraten führen. Die L_1 - oder L_2 -Norm, die zur Berechnung der Ähnlichkeit zweier Histogramme eingesetzt werden, vergleichen übereinstimmende Helligkeitswerte eines Histogramms, so dass sich deutliche Histogrammdifferenzen durch Helligkeitsverschiebungen ergeben können. In der Abbildung 2.7 (a) wird am Beispiel der dargestellten Histogramme deutlich, dass die L_1 - oder L_2 -Norm die Ähnlichkeit zweier Histogramme nur bedingt abbildet. Die Histogrammdifferenzen der L_1 -Norm sind zwischen allen Histogrammen maximal, obwohl sich die ersten beiden Histogramme nur durch eine geringe Verschiebung der Helligkeit unterscheiden.

Die *Earth-Movers*-Distanz bildet Helligkeitsänderungen besser ab [439]. Die Distanz entspricht dem minimalen Aufwand, um ein Histogramm in ein Zweites zu überführen; sie wird aus der Anzahl der zu verschiebenden Pixel und dem Umfang der Verschiebung abgeleitet. Im mehrdimensionalen Fall kann die Berechnung der *Earth-Movers*-Distanz auf ein Transportproblem zurückgeführt werden, wobei durch die Komplexität von $O(n^3 \log n)$ bei einer Anzahl von n Datenelementen die Lösung dieses Problems nur mit hohem Rechenaufwand zu ermitteln ist [202, 440]. Im eindimensionalen Fall, also beim Vergleich zweier Histogramme, entspricht die *Earth-Movers*-Distanz der L_1 -Norm von kumulierten Histogrammen. Für die

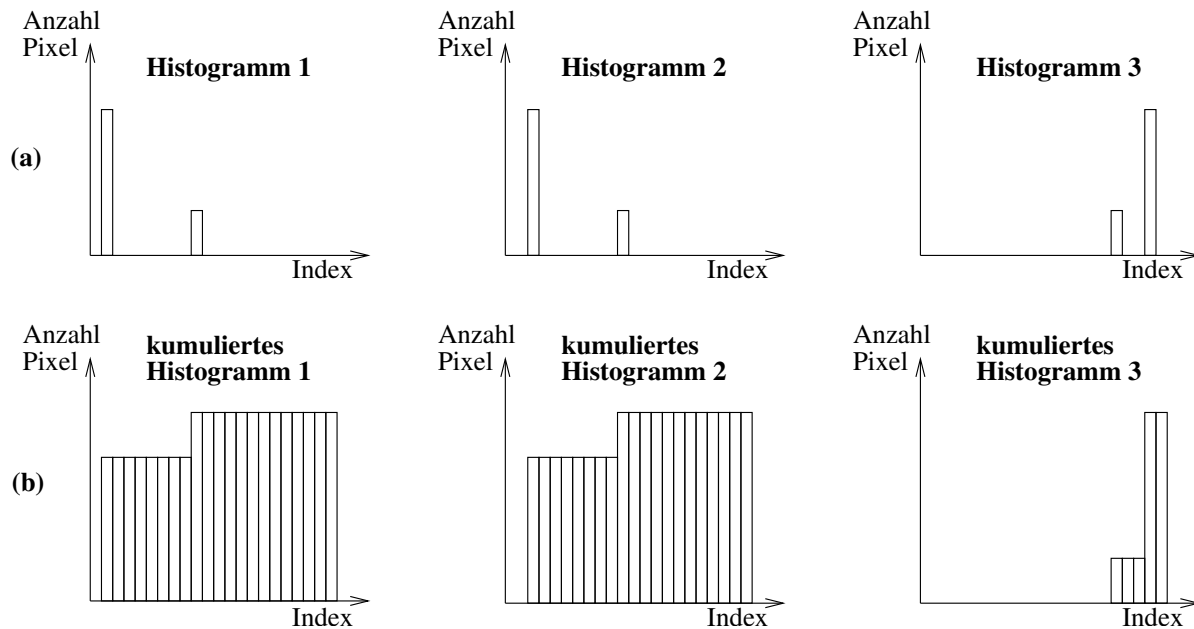


Abbildung 2.7: Vergleich von Histogrammen (a) und kumulierten Histogrammen (b): Die ersten beiden Histogramme unterscheiden sich lediglich durch eine geringe Verschiebung der Helligkeit.

kumulierten Histogramme in der Abbildung 2.7 (b) ist die Histogrammdifferenz der L_1 -Norm zwischen den ersten beiden Histogrammen deutlich niedriger, die Unterschiede zum dritten Histogramm sind jedoch auch für kumulierte Histogramme hoch.

Insbesondere bei Helligkeitsänderungen, die in vielen historischen Videos zwischen benachbarten Bildern zu beobachten sind, bildet die *Earth-Movers*-Distanz die Ähnlichkeit zweier Histogramme deutlich besser ab. Die Ergebnisse der Schnitterkennung bei der Verwendung der *Earth-Movers*-Distanz sind in Tabelle 2.10 dargestellt. Obwohl der F_1 -Wert bei den Farbvideos praktisch unverändert bleibt, ergibt sich für die historischen Videos durch den Wechsel von der L_1 -Norm zur *Earth-Movers*-Distanz eine Verbesserung des F_1 -Wertes um mehr als acht Prozent.

Bei dem zweiten Ansatz zur Verbesserung der Klassifikationsergebnisse für historische Videos werden die einzelnen Verfahren kombiniert, um Fehler möglichst gut auszugleichen. Eine Verbesserung der Ergebnisse ist möglich, da verschiedene Bildfehler unterschiedliche Auswirkungen auf die einzelnen Verfahren haben. In Bildern mit deutlichem Rauschen liefert die Histogrammdifferenz gute Ergebnisse im Vergleich zur Kantenänderungsrate. Andererseits liefert die Kantenänderungsrate bei Helligkeitsschwankungen wesentliche bessere Ergebnisse als der Vergleich mit Histogrammen.

Für den kombinierten Ansatz wird als Klassifikationsverfahren die *Summe der absoluten Differenzen* verwendet, da die Ergebnisse deutlich über allen anderen Verfahren liegen. Die Differenzen bezogen auf die *Kantenänderungsrate* und die *Histogrammdifferenz* werden genutzt, um alle Positionen auszuschließen, an denen mit sehr hoher Wahrscheinlichkeit kein harter Schnitt liegt. Die Schwellwerte der beiden Verfahren werden dabei so festgelegt, dass der Wert für die Vollständigkeit nahe am Maximum von eins liegt. Durch die Kombination der Verfahren verbessert sich der F_1 -Wert um mehr als fünf Prozent auf 91,5 Prozent. Da Bildfehler und Helligkeitsschwankungen in Farbvideos im Allgemeinen nur vereinzelt vorkommen, führt die Kombination der Verfahren zu keiner signifikanten Verbesserung der Ergebnisse.

Trotz einer großen Anzahl an Bildfehlern in den historischen Schwarz-Weiß-Filmen ist durch die Kombination mehrerer Verfahren eine zuverlässige Schnitterkennung möglich. Ergebnisse von mehr als 90 Prozent für die Präzision und Vollständigkeit reichen in vielen Fällen für weitere Analyseverfahren aus. Weiche Schnitte werden in historischen Videos nur vereinzelt eingesetzt und stellen auch in den analysierten Farbvideos nur einen geringen Anteil der Schnitte. Daher hat die höhere Fehlerrate bei der Erkennung weicher Schnitte eine nur geringe Auswirkung auf den Anteil aller fehlerhaft klassifizierten Schnitte.

2.4 Zusammenfassung

In diesem Kapitel wurden Algorithmen zur Erkennung harter und weicher Schnitte analysiert. Zunächst wurde auf die unterschiedlichen Arten von Schnitten eingegangen und die Eigenschaften der Schnitte dargestellt. Anschließend wurde ein dreistufiger Ansatz zur Erkennung von Schnitten vorgestellt, bei dem eine Abbildung vom Bildraum in einen Merkmalsraum definiert, ein Distanzmaß basierend auf den Merkmalen spezifiziert und anhand mehrerer Regeln harte und weiche Schnitte identifiziert wurden.

Im Rahmen der experimentellen Ergebnisse wurden ausgewählte Verfahren zur Erkennung harter und weicher Schnitte analysiert, indem für jeden Ansatz zunächst optimale Schwellwerte für eine Gruppe von Videosequenzen bestimmt wurden und anschließend die Qualität der einzelnen Verfahren anhand einer zweiten Gruppe überprüft wurde.

Abschließend wurde das *European-Chronicles-Online*-Projekt vorgestellt. Die Qualität der innerhalb dieses Projektes analysierten historischen Filme unterscheidet sich deutlich von aktuellen Filmen, so dass mit bestehenden Schnitterkennungsverfahren nur schlechte Klassifikationsergebnisse erzielt wurden. Erst durch die Verwendung der *Earth-Movers*-Distanz wurde für die historischen Videos eine zuverlässige Schnitterkennung möglich. Fehler, die auf Hel-

lichkeitsschwankungen oder Kratzer zurückzuführen sind, konnten zuverlässig durch eine geschickte Kombination der Kantenänderungsrate und Histogrammdifferenzen vermieden werden, so dass auch in den analysierten historischen Videodokumentationen zuverlässige Klassifikationsergebnisse für die Schnitterkennung erreicht wurden.

KAPITEL 3

Analyse der Kamerabewegung

Das zentrale Merkmal eines Videos sind Bildänderungen, die zum größten Teil durch Bewegungen hervorgerufen werden. Es wird zwischen Bewegungen im Bildvordergrund und Bildhintergrund unterschieden, die auch als *Objektbewegungen* (engl. *object motion*) und *Kameraoperationen* bzw. *Kamerabewegungen* (engl. *camera motion*) bezeichnet werden. Anhand der Kamerabewegung können Aussagen über Schnitte abgeleitet oder spezielle Kameraoperationen – wie z. B. Kameraschwenks oder Zoomeffekte – klassifiziert werden. Zusätzlich ist die Kenntnis über die genaue Kamerabewegung Voraussetzung für die bewegungsbasierte Segmentierung von Objekten.

Das zentrale Ziel dieses Kapitels besteht darin, die Kamerabewegung in Videos zu berechnen, um weitere semantische Informationen zu ermitteln. Für die Berechnung der Kamerabewegung wird kurz auf bestehende Verfahren eingegangen und ein geeigneter Ansatz ausgewählt, der effizient zu berechnen ist und präzise Informationen für weitere Analyseschritte liefert. Im Rahmen der experimentellen Ergebnisse wird ausführlich beschrieben, wie ungültige Kameraparameter identifiziert oder eine textuelle Beschreibung der Kamerabewegung ermittelt werden kann. Anschließend wird die Kamerabewegung am Beispiel von Videos in unterschiedlichen Genres analysiert. Aufgrund der charakteristischen Kamerabewegungen ist die Erkennung einzelner Genres wie beispielsweise von Sportveranstaltungen, Zeichentrickfilmen oder Nachrichtensendungen möglich.

Im folgenden Abschnitt wird zunächst ein Modell zur Beschreibung der Kamerabewegung vorgestellt. Abschnitt 3.2 erläutert die Berechnung von Bewegungsvektoren, mit deren Hilfe

die Parameter des Kameramodells in Abschnitt 3.3 geschätzt werden. Um mögliche Ungenauigkeiten der geschätzten Kameraparameter zu reduzieren, wird ein Optimierungsverfahren in Abschnitt 3.4 vorgestellt. Im Rahmen der experimentellen Ergebnisse werden mögliche Fehlerquellen bei der Berechnung des Kameramodells analysiert und ein Verfahren zur Identifikation fehlerhafter Parameter des Kameramodells vorgeschlagen. Zusätzlich werden weitere semantische Informationen über das Video aus den Parametern des Kameramodells abgeleitet.

3.1 Modellierung der Kamerabewegung

Zur Beschreibung der Kamerabewegung zwischen zwei benachbarten Bildern innerhalb einer Kameraeinstellung können verschiedene Modelle eingesetzt werden. Das *zylindrische Kameramodell* (engl. *cylindrical camera model*) [79, 350, 491] projiziert das Bild auf eine Zylinderoberfläche und bildet die horizontale Rotation der Kamera ab, wobei vertikale Rotationen innerhalb des Modells nicht zulässig sind. Das *sphärische Kameramodell* (engl. *spherical camera model*) [91, 492, 553] erweitert das zylindrische Kameramodell, so dass horizontale und vertikale Rotationen möglich werden. Um zusätzlich perspektivische Verzerrungen abzubilden, kann ein Modell mit acht Parametern herangezogen werden [136, 137, 195, 207], das auch im Folgenden verwendet wird. Das Modell bestimmt, ausgehend von der Position (x, y) eines Pixels in Bild i , die neue Position des Pixels (x', y') in Bild $i + 1$:

$$\begin{aligned} x' &= \frac{a_{11}x + a_{12}y + t_x}{p_x x + p_y y + 1}, \\ y' &= \frac{a_{21}x + a_{22}y + t_y}{p_x x + p_y y + 1}. \end{aligned} \quad (3.1)$$

t_x und t_y beschreiben eine horizontale oder vertikale Verschiebung der Bildinhalte, die einem waagrechten oder senkrechten Schwenk (engl. *pan*, *tilt*) der Kamera um den Brennpunkt entspricht. Die Parameter a_{ij} bilden einen Zoomeffekt (engl. *zoom in*, *zoom out*) oder eine Rotation der Kamera entlang der Blickrichtung ab. Die sechs Parametern t_x, t_y und $a_{i,j}$ beschreiben eine affine Transformation [183, 536]. Durch die unterschiedlichen Entfernungen der sichtbaren Objekte zur Kamera können bei einer Drehung der Kamera um den Brennpunkt Verzerrungen auftreten, die durch die beiden Parameter p_x und p_y beschrieben werden. Kamerafahrten (engl. *dolly shot*) werden durch das Modell nicht abgebildet, da keine dreidimensionalen Informationen über die Objekte des Bildes zur Verfügung stehen.

Für die Berechnung des Kameramodells, welche die Änderungen des Bildhintergrundes zwischen zwei benachbarten Bildern beschreiben, wird ein dreistufiges Verfahren verwendet [136]. Zunächst werden Bewegungsvektoren bestimmt, um die Verschiebung einzelner Pixel zwischen den beiden Bildern zu beschreiben. Zur Berechnung der acht Parameter des Kameramodells reicht es aus, die genaue Verschiebung von vier Pixeln des Bildhintergrundes zwischen beiden Bildern zu kennen. Aus den berechneten Bewegungsvektoren werden die Modellparameter geschätzt, so dass das Kameramodell die gefundenen Bewegungsvektoren möglichst gut annähert. Es können geringe Ungenauigkeiten bei der Schätzung der Parameter auftreten, da die Pixelverschiebungen durch ganzzahlige Werte beschrieben werden. In einem dritten Schritt wird daher die Genauigkeit der Modellparameter durch ein Gradientenabstiegsverfahren verbessert, so dass der Unterschied zwischen dem ersten mit dem Kameramodell transformierten Bild und dem zweiten Bild minimal wird. In den folgenden drei Abschnitten werden die einzelnen Schritte des Verfahrens kurz erläutert.

3.2 Berechnung von Bewegungsvektoren

Bewegungen zwischen zwei Bildern eines Videos können durch Bewegungsvektoren (engl. *motion vector*) beschrieben werden. Wird für jedes Pixel des Bildes ein Bewegungsvektor bestimmt, der die Verschiebung des Pixels vom ersten zum zweiten Bild beschreibt, so spricht man vom *optischen Fluss* (engl. *optical flow*) [27, 205, 206, 473]. Eine große Anzahl an Verfahren zur optimierten Berechnung des optischen Flusses wurden entwickelt [25, 37, 127, 460, 538, 539, 568].

Statt ein dichtes Feld mit Vektoren zu bestimmen, reicht es zur Berechnung der Kamerabewegung aus, Bewegungsvektoren für Bildbereiche oder einzelne im Bild verstreute Merkmale (engl. *sparse features*) zu berechnen [3]. Zur Berechnung der Bewegungsvektoren werden eindeutige hervorstechende Merkmale (engl. *salient points*) im ersten Bild identifiziert und die entsprechenden Merkmale im zweiten Bild gesucht [329, 355]. *Ecken von Bildregionen* eignen sich durch ihre eindeutige Struktur besonders gut als Merkmal. Wir haben das nach Harris [194] benannte Verfahren zur Erkennung von Ecken in Bildern ausgewählt, da es auch bei starkem Rauschen und feinen Texturen sehr zuverlässige Ergebnisse liefert [134, 456].

Nach der Ermittlung der signifikanten Ecken in den beiden Bildern i und j müssen diese einander paarweise zugeordnet werden [341]. Die Positionen jeweils zweier zugeordneter Ecken definieren einen Bewegungsvektor. Die grundlegende Idee ist auf das Verfahren von Zhang et al. zurückzuführen, bei dem die Blöcke um jeden Merkmalspunkt analysiert und die Korre-

lationen zwischen Blöcken der beiden Bilder berechnet werden [586]. Zunächst werden alle möglichen Kombinationen der Ecken aus Bild i und j betrachtet. Jede Kombination wird als ein möglicher Bewegungsvektor interpretiert, für den als Qualitätsmaß die Summe der absoluten Differenzen der umgebenden Blöcke berechnet wird. Liegen die Positionen (x, y) und (x', y') der Ecken sehr weit auseinander, so wird angenommen, dass es sich um keine gültige Zuordnung handelt. Insbesondere bei Kameraschwenks wird so verhindert, dass Ecken, die im ersten Bild noch vorhanden waren, aber im zweiten Bild nicht mehr sichtbar sind, einer falschen Ecke zugewiesen werden.

Die Zuordnung der Ecken erfolgt durch einen *Greedy*-Algorithmus [19, 92]. Die beiden ähnlichsten Ecken in Bezug auf die Summe der absoluten Differenzen der umgebenden Blöcke werden einander zugeordnet, als ausgewählt markiert und definieren einen Bewegungsvektor. Iterativ werden weitere Bewegungsvektoren erzeugt, indem unter den verbleibenden Ecken die jeweils ähnlichsten kombiniert und markiert werden. Das Verfahren terminiert, wenn die Differenzen einen Schwellwert übersteigen und angenommen werden kann, dass keine korrekte Zuordnung von Ecken mehr möglich ist. Die Zuordnung von Ecken ermöglicht eine schnelle, zuverlässige und auch bei starken Kamerabewegungen verwendbare Berechnungsmethode zur Ermittlung von Bewegungsvektoren. Ungenauigkeiten können auftreten, wenn durch Rauschen und Objektbewegungen Ecken in Videos verschwinden und neue Ecken erscheinen.

Abbildung 3.1 (a) und (b) zeigt zwei Bilder einer Kameraeinstellung mit einem horizontalen Schwenk. Zum Vergleich sind die Bewegungsvektoren, die bei einer Verwendung des *Block-matching*-Verfahrens entstehen würden [536], in (c) abgebildet, wobei insbesondere im Bereich des Himmels, der keine eindeutigen Strukturen aufweist, deutliche Fehler bei den Vektoren auftreten. Die signifikanten Ecken der beiden Bilder sind in (d) und (e) markiert, aus denen durch Zuordnung Bewegungsvektoren abgeleitet werden (f). Die Qualität dieser Bewegungsvektoren ist sehr unterschiedlich. Ein hoher Anteil der Bewegungsvektoren beschreibt die Veränderung der Bildpositionen sehr genau, wobei insbesondere durch die Bewegungen der Personen im unteren Bildbereich deutliche Unterschiede zu den erwarteten Bewegungsvektoren des Kameramodells auftreten.

3.3 Schätzung der Parameter des Kameramodells

Um aus den Bewegungsvektoren die Parameter des Kameramodells zu berechnen, muss erkannt werden, ob ein Vektor die Bewegung des Bildhintergrundes korrekt beschreibt. Eine fehlerhafte Länge oder Richtung eines Vektors entsteht häufig bei Objektbewegungen im Bild-

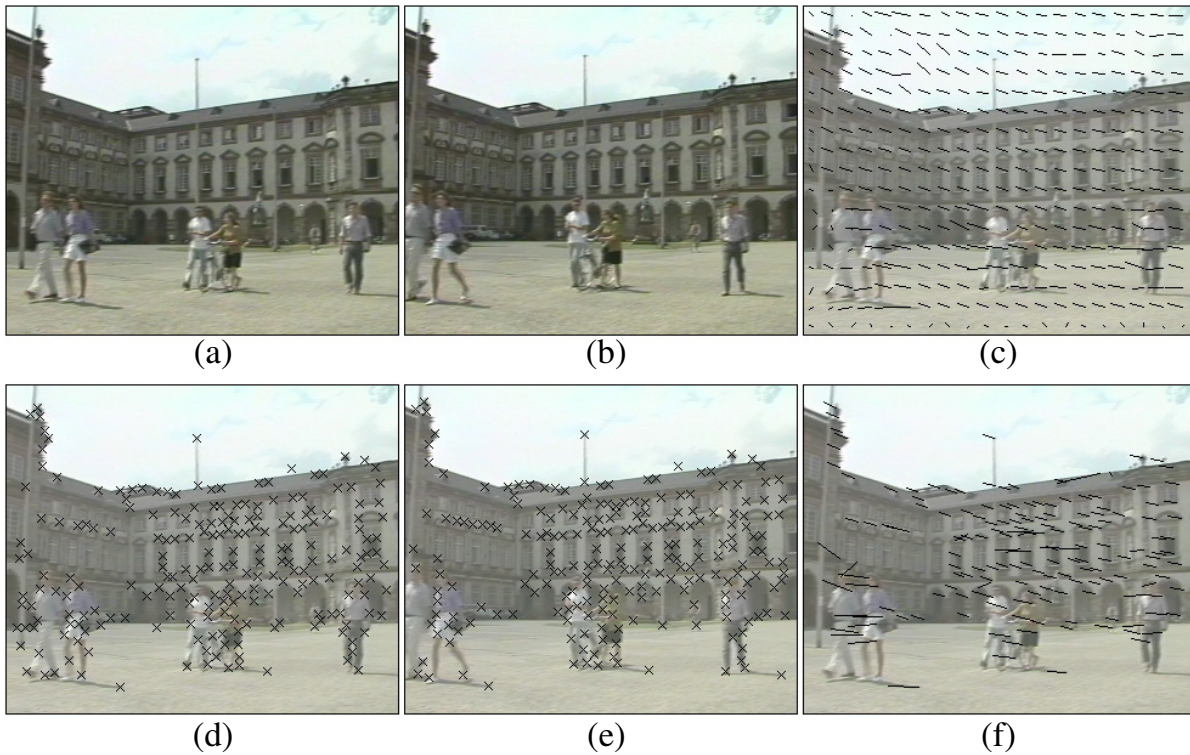


Abbildung 3.1: Schätzung der Bewegungsvektoren für zwei Bilder (a) und (b) einer Videosequenz : Bewegungsvektoren des Blockmatching-Verfahrens (c), signifikante Ecken der beiden Bilder (d) und (e), aus den Ecken abgeleitete Bewegungsvektoren (f).

vordergrund. Vor der Berechnung des Kameramodells ist jedoch nicht bekannt, ob ein Vektor Bewegungen im Vorder- oder Hintergrund beschreibt. Das in diesem Abschnitt verwendete Verfahren [501] berechnet iterativ die Parameter des Kameramodells anhand weniger Bewegungsvektoren und prüft, wie gut das Kameramodell mit allen Bewegungsvektoren übereinstimmt.

Unter der Annahme, dass mindestens die Hälfte der Vektoren die Bewegung des Hintergrundes beschreibt, kann die Kamerabewegung mit einer robusten Regressionsschätzung berechnet werden. Ausgewählt wurde das Verfahren der *kleinsten getrimmten Quadrate* (engl. *least trimmed squares*) [432, 433], bei dem zunächst zufällig vier Bewegungsvektoren aus der Menge aller Vektoren ausgewählt werden. Mit diesen vier Vektoren ist es möglich, durch Lösen eines linearen Gleichungssystems mit acht Gleichungen die acht Parameter des Kameramodells eindeutig zu berechnen. Jeder Bewegungsvektor beschreibt eine Positionsänderung in horizontaler und vertikaler Richtung, so dass insgesamt acht Wertepaare zum Lösen des Gleichungs-

systems zur Verfügung stehen.

Eine Fehlerfunktion klassifiziert den Fehler zwischen der tatsächlichen Position eines Pixels (x', y') im zweiten Bild und der durch das Kameramodell geschätzten Position (\hat{x}, \hat{y}) . Für jeden Bewegungsvektor i ($i = 1 \dots N$) wird ein Fehler e_i anhand der quadrierten euklidischen Distanz berechnet:

$$e_i = (x'_i - \hat{x}_i)^2 + (y'_i - \hat{y}_i)^2. \quad (3.2)$$

Da Vektoren aus Bereichen des Bildvordergrundes die Fehlerfunktion nicht beeinflussen sollen, wird nur der Teil der Vektoren betrachtet, der gut zum Modell passt. Die Fehler der einzelnen Vektoren werden aufsteigend nach ihrer Größe sortiert, so dass in der zweiten Hälfte der Liste die Vektoren enthalten sind, die stärker vom Kameramodell abweichen. Die Hälfte mit den geringeren Fehlerwerten wird zum gesamten Fehler E aufsummiert:

$$E = \sum_{i=1}^{N/2} e_i \quad \text{mit } e_1 \leq \dots \leq e_N. \quad (3.3)$$

Um den Fehler E zu minimieren, wird das Verfahren der kleinsten getrimmten Quadrate mehrfach angewendet. Iterativ werden jeweils vier Bewegungsvektoren aus der Menge aller Bewegungsvektoren zufällig ausgewählt, für die die Parameter des Kameramodells, die Fehler der Bewegungsvektoren und der gesamte Fehler E berechnet werden. Die Parameter des Kameramodells mit dem minimalen Fehler werden gespeichert.

Mit der Anzahl der Iterationen steigt die Wahrscheinlichkeit, dass mindestens einmal vier Bewegungsvektoren zufällig ausgewählt werden, welche die Bewegung des Bildhintergrundes gut beschreiben. Bei Vorgabe einer gewünschten Wahrscheinlichkeit kann die Anzahl der notwendigen Iterationen genau bestimmt werden. Das Verfahren liefert zuverlässige Ergebnisse, solange mindestens die Hälfte der Bewegungsvektoren die Bewegung des Bildhintergrundes beschreibt. Durch das Verwerfen aller stark vom Kameramodell abweichenden Bewegungsvektoren ist auch bei fehlerhaften Bewegungsvektoren und Objektbewegungen eine zuverlässige Berechnung der Kamerabewegung möglich.

Bei den hell markierten Bewegungsvektoren in Abbildung 3.2 (links) handelt es sich um die Vektoren, die stark vom Kameramodell abweichen und keinen Einfluss auf die Parameter des Kameramodells haben. Die Ähnlichkeiten der dunkel markierten Bewegungsvektoren im lin-



Abbildung 3.2: Links: Die hell markierten Bewegungsvektoren weichen von der Bewegung des Bildhintergrundes deutlich ab, die dunkel markierten Vektoren werden zur Berechnung des Kameramodells verwendet.
Rechts: Bewegungsvektoren des automatisch berechneten Kameramodells.

ken Bild mit den Vektoren des automatisch berechneten Kameramodells im rechten Bild sind sehr groß. Im rechten Bild weichen die Bewegungsvektoren im unteren linken bzw. oberen rechten Bildbereich von der tatsächlichen Kamerabewegung ab, da keine geeigneten Bewegungsvektoren in diesen Bereichen erkannt werden.

3.4 Exakte Berechnung des Kameramodells

Die Genauigkeit der aus den vier Bewegungsvektoren berechneten Parameter des Kameramodells reicht für weitere Analyseschritte nicht immer aus. Dies ist im Wesentlichen auf die ungenauen Positionen der erkannten Ecken zurückzuführen, die lediglich pixelgenau bestimmt werden, so dass auch die Bewegungsvektoren nur ganzzahlige Werte annehmen können. Eine ganzzahlige Verschiebung von Pixeln entspricht jedoch nicht der Realität und führt zu ungenauen Parametern des Kameramodells. Obwohl die Abweichung der tatsächlichen Parameter von den ermittelten Parametern nicht sehr groß ist, reicht die verfügbare Genauigkeit insbesondere für die Erzeugung von Panoramabildern oder zur bewegungsbasierten Segmentierung von Objekten nicht aus.

Zur Verbesserung der Genauigkeit des Kameramodells wird auf das von Irani et al. entwickelte Verfahren zurückgegriffen [223]. Dabei werden die vorhandenen Fehler der Kameraparameter

durch Minimierung der Differenz zwischen dem mit dem Kameramodell transformierten Bild \hat{I}_i und dem zweiten Bild I_j verringert. Die Differenz $E_{i,j}$ der beiden Bilder i und j ist definiert als:

$$E_{i,j} = \sum_{x,y} \begin{cases} e(x,y) & \text{falls } e(x,y) < t, \\ t & \text{sonst,} \end{cases} \quad (3.4)$$

mit $e(x,y) = (\hat{I}_i(x,y) - I_j(x,y))^2$.

$\hat{I}_i(x,y)$ ist die Helligkeit des Pixels an der Position (x,y) im transformierten ersten Bild, $I_j(x,y)$ der Helligkeitswert im zweiten Bild. Der maximale Fehler eines Pixels ist durch einen Schwellwert t nach oben beschränkt, um den Einfluss von Objektbewegungen zu reduzieren. Ansonsten würden sich beim Optimierungsprozess die Parameter des Kameramodells so anpassen, dass neben dem Bildhintergrund auch der Bildvordergrund möglichst deckungsgleich wird.

Das *Gradientenabstiegsverfahren* (engl. *gradient descent*) [21, 33, 223, 297] eignet sich als heuristisches Verfahren zur Berechnung eines lokalen Minimums für den Fehler. Eine erste grobe Schätzung der Parameter p^0 des Kameramodells ist aus dem vorherigen Schritt bekannt. Die folgende jeweils verbesserte Schätzung p^{n+1} wird berechnet durch:

$$p^{n+1} = p^n - \alpha \cdot \nabla E^n. \quad (3.5)$$

α ist eine Konstante und gibt die Schrittweite für jede Iteration an. ∇ bezeichnet den Gradienten der Fehlerfunktion E , die minimiert werden soll. Mit jeder Iteration werden die Parameter des Kameramodell so angepasst, dass sich der Fehler E verringert. Das Verfahren terminiert, falls die Änderungen zwischen zwei Iterationen sehr gering werden und keine deutliche Reduktion des Fehlers mehr möglich ist. Es gibt eine Vielzahl effizienter und stabiler Algorithmen, die das Gradientenabstiegsverfahren nutzen und ein schnelles Konvergieren zum lokalen Minimum gewährleisten [305, 346, 423].

Durch den Optimierungsschritt steigt die Genauigkeit der Parameter des Kameramodells signifikant. Zur Berechnung des Fehlers werden nicht nur einzelne Bewegungsvektoren verwendet, sondern alle Pixel eines Bildes. Da bei der Transformation des Bildes die Helligkeitswerte der Pixel durch Nachbapixel interpoliert werden, erhöht das Verfahren die Präzision der Parame-

ter des Kameramodells auf Subpixelgenauigkeit. Diese hohe Genauigkeit ist insbesondere für die Erzeugung von Panoramabildern oder zur Objektsegmentierung mittels Bewegungsanalyse erforderlich.

Die Kombination der beiden Verfahren, also die Schätzung der Bewegungsvektoren durch Zuordnung der Ecken und die exakte Berechnung der Kameraparameter mit dem Gradientenabstiegsverfahren, ermöglicht eine zuverlässige Berechnung des Kameramodells. Eine schnelle Schätzung der Bewegungsvektoren ist auch bei starken Kamerabewegungen möglich, und das Gradientenabstiegsverfahren liefert ausgehend von der ersten Schätzung sehr präzise Parameter des Kameramodells.

3.5 Experimentelle Ergebnisse

Bei der Berechnung der Kameraparameter können in jedem einzelnen Schritt Fehler auftreten, die in den nachfolgenden Schritten nicht mehr korrigierbar sind. In Kameraeinstellungen mit niedrigem Kontrast ist die Anzahl der erkannten Ecken und gültigen Bewegungsvektoren möglicherweise so gering, dass eine Berechnung der Parameter des Kameramodells nicht möglich ist. Weiterhin führen regelmäßige Strukturen im Bild zu einer hohen Anzahl fehlerhafter Bewegungsvektoren.

Große Objekte im Bildvordergrund können die Berechnung der Parameter des Kameramodells verhindern. Übersteigt die Anzahl der Bewegungsvektoren im Bildvordergrund die des Bildhintergrundes, so werden auch die Parameter des Kameramodells die Bewegungen der Objekte im Vordergrund beschreiben. Bei einer großen Abweichung der geschätzten von den tatsächlichen Parametern des Kameramodells kann auch das Gradientenabstiegsverfahren die Ergebnisse nicht verbessern, da ausgehend von der ersten Schätzung ein lokales und nicht das globale Minimum für den Fehler gesucht wird.

Ein großer Anteil der möglichen Fehler kann automatisch erkannt werden. Bei geringem Kontrast ist die Anzahl der erkannten Ecken sehr gering, und eine Berechnung des Kameramodells ist nicht möglich. Sind ausreichend viele Bewegungsvektoren verfügbar, so werden die Parameter des Kameramodells geschätzt. Der Unterschied zwischen den geschätzten und tatsächlichen Bewegungsvektoren liefert einen Hinweis auf die Qualität der ermittelten Parameter. Übersteigt der Fehler einen Schwellwert, so handelt es sich offensichtlich um falsche Parameter des Kameramodells.

In einem letzten Schritt wird überprüft, ob die Parameter des Kameramodells gültige Werte annehmen. Das Modell zur Beschreibung der Kamerabewegung (vgl. Gleichung 3.1) bildet

Kameraoperation	t_x, t_y	a_{11}, a_{22}	a_{12}, a_{21}	p_x, p_y
Statische Kamera	0	1	0	0
Translation	$\neq 0$	1	0	0
Skalierung				
- Zoom-in	0	$0 < a_{00} = a_{11} < 1$	0	0
- Zoom-out	0	$a_{00} = a_{11} > 1$	0	0
Rotation um Winkel θ	0	$a_{00} = a_{11} = \cos \theta$	$a_{01} = -a_{10} = \sin \theta$	0
Scherung				
- horizontal	0	1	$a_{01} \neq 0$	0
- vertikal	0	1	$a_{10} \neq 0$	0
Spiegelung				
- horizontal	0	$a_{00} = -1$	0	0
- vertikal	0	$a_{11} = -1$	0	0
Persp. Verzerrung	0	1	0	$\neq 0$

Tabelle 3.1: Auswirkung einer Kameraoperation auf die Parametern des Kameramodells

affine Transformationen und perspektivische Verzerrungen ab. Nur ein kleiner Teil der durch das Modell abbildbaren Transformationen kann in Kameraeinstellungen tatsächlich beobachtet werden [47, 556].

Tabelle 3.1 verdeutlicht den Zusammenhang zwischen den Parametern des Kameramodells und den entsprechenden Transformationen. Abgesehen von Kameraeinstellungen mit statischer Kamera sind Kameraschwenks und Zoom-Effekte die mit Abstand am häufigsten auftretenden Kameraoperationen in Videos. Bei unterschiedlicher Entfernung der sichtbaren Objekte zur Kamera sind bei Kameraschwenks perspektivische Verzerrungen möglich. Eine Rotation der Kamera ist zwar denkbar, wird jedoch nur extrem selten eingesetzt. Lediglich bei verwackelten Aufnahmen ohne Stativ können kurzzeitig schwache Rotationen auftreten. Der Rotationswinkel θ ist in diesem Fall sehr gering, und die Rotationsrichtung wechselt innerhalb weniger Bilder. Obwohl das Kameramodell Spiegelungen und Scherungen abbildet, können diese in realen Kameraeinstellungen nicht vorkommen.

Abbildung 3.3 verdeutlicht exemplarisch die Veränderung der Bildinhalte in Abhängigkeit der Parameter des Kameramodells. Bis auf die Parameter t_x und t_y , die eine Translation beschreiben, sind die Abweichungen der Parameter vom Mittelwert sehr gering. Bei sehr schnellen und starken Kameraoperationen können die Parameter $a_{i,j}$ zweier benachbarter Bilder einer Kameraeinstellung um bis zu 0,1 vom Mittelwert abweichen. Die Parameter, die die perspektivische Verzerrung beschreiben, liegen sehr nahe bei null, und kleinste Abweichungen verursachen, wie in Abbildung 3.3 (g) deutlich zu sehen ist, signifikante Verzerrungen des Bildes.

In Videos müssen die Werte der Kameraparameter innerhalb fester Intervalle liegen, damit sie

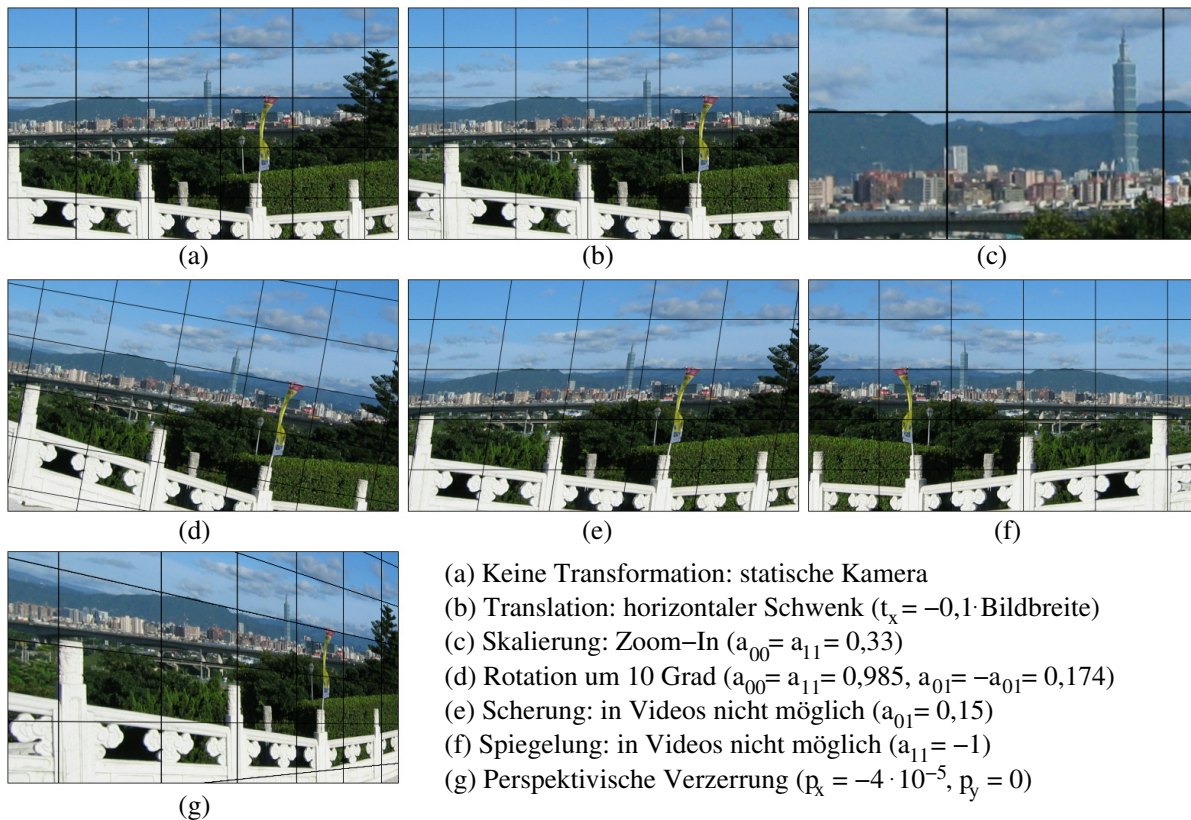


Abbildung 3.3: Bildänderungen bei unterschiedlichen Transformationen

eine reale Kamerabewegungen beschreiben. Zulässige Werte für die unterschiedlichen Kameraoperationen sind in Tabelle 3.2 aufgelistet. Lediglich die Parameter t_x und t_y nehmen bei starken Kameraschwenks höhere Werte an, deren maximaler Wert auf $\frac{1}{5}$ der Bildhöhe H bzw. der Bildbreite W beschränkt wird. Diese schnellen Schwenks, bei denen sich innerhalb eines Bruchteils einer Sekunde der Bildinhalt vollständig ändert, werden als *Reißschwenk* (engl. *swish pan*) bezeichnet und häufig in Kombination mit Schnitten eingesetzt.

Die Intervalle in Tabelle 3.2 wurden so festgelegt, dass sich während eines Zoomeffektes die Größe der Objekte im Zentrum des Bildes um maximal acht Prozent zwischen zwei Bildern ändert. In einer Videosequenz mit 25 Bildern pro Sekunde ist bei diesem Wert eine theoretisch maximale Vergrößerung um den Faktor acht innerhalb einer Sekunde möglich. Eine Rotation der Kamera entlang der Blickrichtung kommt sehr selten vor, und lediglich bei Aufnahmen ohne Stativ kann eine geringe Neigung der Kamera beobachtet werden. Eine Rotation der Kamera um bis zu fünf Grad ist mit den in Tabelle 3.2 angegebenen Parametern zulässig. Die Parameter, die die perspektivische Verzerrung beschreiben, weichen nur minimal von null ab

Kameraoperation	t_x, t_y	a_{11}, a_{22}	a_{12}, a_{21}	p_x, p_y
statische Kamera	$0 \pm 0,8$	$1 \pm 0,01$	$0 \pm 0,01$	$0 \pm 1 \cdot 10^{-6}$
horizontaler Kameraschwenk	$0 \pm \frac{1}{5}W$	$1 \pm 0,02$	$0 \pm 0,02$	$0 \pm 2 \cdot 10^{-4}$
Zoomeffekt	$0 \pm 0,8$	$1 \pm 0,08$	$0 \pm 0,08$	$0 \pm 1 \cdot 10^{-5}$
Rotation (max. $\theta = 5^\circ$)	$0 \pm 0,8$	$1 \pm 0,01$	$0 \pm 0,09$	$0 \pm 1 \cdot 10^{-5}$

Tabelle 3.2: Gültige Intervalle für die acht Parameter des Kameramodells bei unterschiedlichen Kameraoperationen. W definiert die Bildbreite.

und wurden experimentell ermittelt.

Neben der Überprüfung, ob es sich um plausible Parameter des Kameramodells handelt, kann aus den Werten der Parameter eine Beschreibung der Kameraoperation abgeleitet werden. Beispielsweise ist es möglich, *Start*, *Länge* und *Stärke* eines Kameraschwenks oder Zoomeffektes automatisch zu charakterisieren. Anhand der Rotation können Rückschlüsse über die Art der Aufnahme gezogen werden, so dass beispielsweise erkannt werden kann, ob ein *Stativ* bei der Filmaufnahme verwendet wurde.

Abbildung 3.4 verdeutlicht die Änderung der Parameter des Kameramodells innerhalb einer Videosequenz. Durch die Analyse der Parameter kann automatisch erkannt werden, dass in der ersten Kameraeinstellung (bis einschließlich Bild 124) ohne Stativ gefilmt wurde und ein eingehender Zoomeffekt vorkommt. Starke und kurzfristige Schwankungen der geglätteten Werte t_x , t_y und a_{10} deuten auf eine verwackelte Kameraführung hin. Die negativen Werte von a_{00} über einen Zeitraum von mehreren Sekunden ermöglichen die automatische Erkennung des Kamerazooms. In der zweiten, mit einem Stativ aufgenommenen Kameraeinstellung (ab Bild 125) tritt – deutlich erkennbar an den Werten des Parameters t_x – zunächst ein horizontaler Schwenk auf. Die Kameraeinstellung geht in eine Aufnahme mit einer statischen Kamera über. Am Beispiel der neun in Kapitel 2.3 vorgestellten Testsequenzen wird analysiert, wie präzise die Parameter des Kameramodells ermittelt werden. Anhand der aggregierten Ergebnisse in Tabelle 3.3 wird deutlich, dass mehr als 94 Prozent der Kameraparameter korrekt berechnet und die Parameter des Kameramodells sehr zuverlässig bestimmt werden können. Fehlerhafte Parameter treten verstärkt bei offenem Feuer, sich ändernden Lichtverhältnissen, bei großen sich bewegenden Objekten und bei harten und weichen Schnitten auf. Tabelle 3.3 gibt den Anteil und die wesentlichen Ursachen für die beobachteten Fehler an.

In Tabelle 3.4 ist die Anzahl der erkannten Kameraoperationen für alle Testsequenzen angegeben, wobei nur die deutlich ausgeprägten Kameraoperationen aufgeführt sind. Sehr schwache oder kurze Schwenks und Zoomeffekte, wie sie beispielsweise in verwackelten Kameraauf-

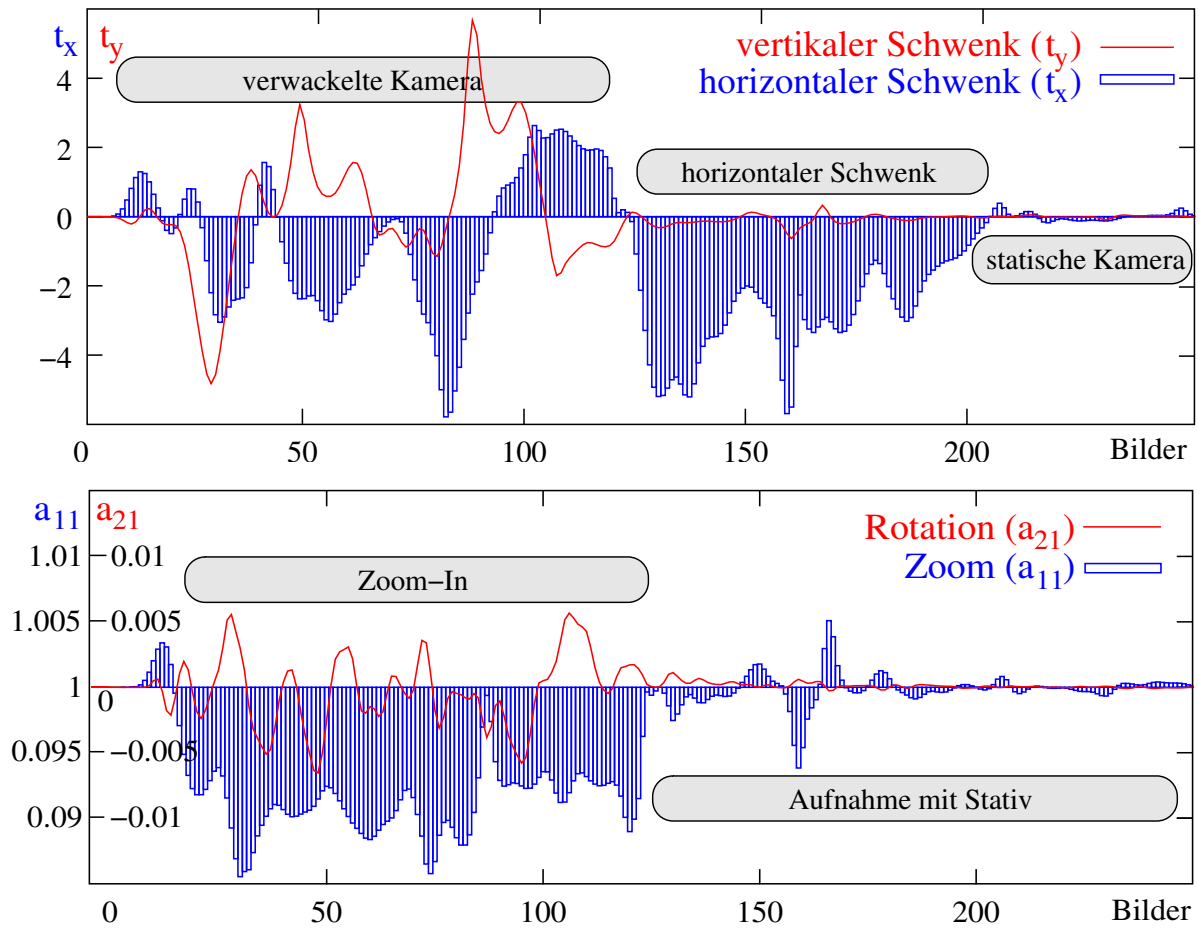


Abbildung 3.4: Klassifikation einer Kameraoperation durch Analyse der Parameter des Kameramodells. Ein Kameraschwenk, eine Zoomoperation und die Verwendung eines Stativs können automatisch erkannt werden.

nahmen auftreten, sind nicht in den Ergebnissen enthalten. Damit eine Kameraoperation als Schwenk erkannt wird, müssen die absoluten Werte der Parametern t_x oder t_y für mindestens 1,5 Sekunden deutlich von null abweichen. Da die Dauer der Zoomeffekte häufig geringer ist, wurde für diese eine Mindestlänge von einer Sekunde spezifiziert.

Bei der Analyse der Kameraoperationen der Testsequenzen fällt auf, dass deutlich mehr horizontale als vertikale Schwenks auftreten. Ähnliches gilt für die Zoomoperationen, bei denen die Anzahl der ausgehenden Operationen die der eingehenden Zoomoperationen deutlich übersteigt. Häufig befindet sich zu Beginn einer Zoomoperation das im Zentrum des Interesses liegende Objekt nicht in der Bildmitte, so dass zusätzlich zum Zoom ein Kameraschwenk beobachtet wird.

Innerhalb einer Nachrichtensendung liegt der Anteil der Kameraeinstellungen, in denen der

Kameramodell	Anteil	Ursache
korrekt erkannt	94,8 %	
Ecken wurden nicht erkannt	0,3 %	geringer Kontrast
Zuordnung der Ecken zu den Bewegungsvektoren nicht möglich	0,1 %	harte Schnitte, plötzliche Bildänderungen
fehlerhaftes Modell	4,8 %	harte und weiche Schnitte, Objektbewegungen

Tabelle 3.3: Anteil der Bilder mit korrekt und fehlerhaft berechneten Parametern des Kameramodells für die neun Testsequenzen

Sprecher zu sehen ist, bei ungefähr dreißig Prozent. Im Gegensatz zu diesen nahezu statischen Aufnahmen ist der Anteil der Kameraoperationen in den Beiträgen einer Nachrichtensendung überproportional hoch. Umgekehrt sieht die Situation bei dem analysierten Zeichentrickfilm aus, in dem nur sehr vereinzelt Kamerabewegungen beobachtet werden können. Ein charakteristisches Merkmal von Sportsendungen ist eine große Anzahl schneller horizontaler Schwenks, da die Kamera dem aktuellen Spielgeschehen folgt. Obwohl in Werbefilmen deutliche Bewegungen auftreten, ist der Anteil der Kameraoperationen sehr gering. Das liegt im Wesentlichen an der vorgegebenen Mindestlänge einer Kameraoperation und der hohen Anzahl an Schnitten in Werbevideos. Weiterhin ist das Verhältnis der eingehenden zu den ausgehenden Zoomeffekten überproportional hoch, da relativ häufig Produkte oder Markennamen durch eine eingehende Zoomoperation hervorgehoben werden.

Zum Teil ist es möglich, aus den automatisch ermittelten Kameraoperationen das Genre des Videos zu ermitteln. In Nachrichtensendungen wechseln sich lange statische mit kurzen dynamischen Kameraeinstellungen ab, Sportsendungen enthalten viele horizontale Schwenks, und der Anteil der Kameraoperationen in Zeichentrickfilmen ist sehr gering. Die Erkennung ist jedoch nur für ausgewählte Genres möglich, da nicht in jedem Genre charakteristische Kameraoperationen verwendet werden.

3.6 Zusammenfassung

In diesem Kapitel wurde ein Verfahren vorgestellt, um die Kamerabewegung in Videos zu ermitteln, so dass diese zur Berechnung weiterer semantischer Informationen genutzt werden kann. Hierbei wurde auf ein bekanntes Verfahren zur Schätzung der Kameraparameter durch Zuordnung der Ecken zweier Bilder zurückgegriffen. Der Optimierungsschritt mit dem Gradientenabstiegsverfahren führte zu einer sehr genauen und zuverlässigen Berechnung der

	horizontaler Schwenk	vertikaler Schwenk	eingehender Zoom	ausgehender Zoom
Dokumentation	31	12	12	21
Nachrichtensendung	40	18	14	30
Spielfilm	32	4	15	33
Talkshow	41	9	28	48
Serie	18	11	19	24
Zeichentrickfilm	3	1	2	16
Sportsendung	81	7	13	28
Musikclip	27	10	10	24
Werbung	18	19	18	20
Summe	301	88	123	254

Tabelle 3.4: Anzahl der automatisch erkannten Kameraoperationen in den Testsequenzen

Parameter des Kameramodells.

Im Rahmen der experimentellen Ergebnisse wurde ein Verfahren zur Erkennung ungültiger Kameraparameter vorgestellt. Zudem wurde eine textuelle Beschreibung der Kamerabewegung aus den Kameraparametern abgeleitet. Durch die Analyse der charakteristischen Kamerabewegungen konnte das Genre eines Videos beispielsweise für Sportveranstaltungen, Nachrichtensendungen oder Zeichentrickfilme zuverlässig bestimmt werden.

Zusammenfassend bleibt festzuhalten, dass die Analyse der Bewegung wichtige Informationen über ein Video wie beispielsweise die Länge und Stärke der verwendeten Kameraoperationen, die Rückschlüsse auf das Genre des Videos zulassen, liefert. Ob ein Video mit oder ohne Stativ aufgenommen wurde, kann ebenfalls erkannt werden. Weiterhin ist die Kamerabewegung Voraussetzung für die bewegungsbasierte Objektsegmentierung, auf die in Kapitel 4 eingegangen wird.

KAPITEL 4

Objektsegmentierung durch Bewegungsanalyse

In diesem Kapitel wird ein Verfahren vorgestellt, um Objekte des Bildvordergrundes, d. h. Objekte, deren Bewegungen sich von der des Bildhintergrundes unterscheiden, zu segmentieren. Dieser Schritt ist Voraussetzung für die Objekterkennung in Videos und liefert Informationen über die genauen Positionen und Formen der Objekte im Bild. Die Segmentierung schafft die Möglichkeit zur nachträglichen Änderung von Filmen, indem Objekte ausgeschnitten und neue Objekte in einen Film eingesetzt werden können. Während der Segmentierung werden Hintergrund- bzw. Panoramabilder erzeugt, welche die Grundlage für bildbasierte Zusammenfassungen von Videos liefern.

Ziel dieses Kapitels ist es nicht, das Problem der Objektsegmentierung in voller Breite zu behandeln. Das vorgestellte Verfahren zur Objektsegmentierung ist vielmehr Voraussetzung für die Objekterkennung im folgenden Kapitel. Als wesentliche Ideen werden in diesem Kapitel ein neues Verfahren zur zuverlässigen Segmentierung bei langsamen Objektbewegungen, ein neuer Algorithmus zur Analyse der Randbereiche der segmentierten Objekte sowie ein neuer Ansatz zur Transformation von Farbbildern vorgestellt.

Zur Identifikation der Objektgrenzen wird die Bewegung des Objektes mit der Bewegung der Kamera verglichen. Stoppt die Objektbewegung innerhalb der Kameraeinstellung, so ist eine zuverlässige Erkennung des Objektes nicht mehr möglich. Mit dem in diesem Kapitel vorgestellten Verfahren können beliebig viele Objekte des Bildes gleichzeitig segmentiert werden, solange ein deutlicher Helligkeits- oder Farbunterschied zwischen Objekt und Hintergrund

besteht, mindestens die Hälfte der Pixel in jedem Bild zum Bildhintergrund gehören und kontinuierliche Objektbewegungen auftreten.

Die Segmentierung eines Objektes erfolgt in drei Schritten. Zunächst wird in den Abschnitten 4.1 und 4.2 ein Verfahren vorgestellt, um den Bildhintergrund in allen Bildern der Kameraeinstellung deckungsgleich auszurichten. Dabei werden aus den bekannten Parametern des Kameramodells von jeweils zwei aufeinander folgenden Bildern die Modellparameter zwischen beliebigen Bildern der Kameraeinstellung hergeleitet. Zwei Verfahren zur Transformation eines Bildes werden erläutert, wobei das erste Verfahren besonders gut zur Objektsegmentierung geeignet ist und das zweite Verfahren Vorteile für die Erzeugung von Panoramabildern bietet. In einem zweiten Schritt wird in Abschnitt 4.3 die Erzeugung eines Hintergrundbildes vorgestellt, in dem Vordergrundobjekte nicht mehr enthalten sind. Eine besondere Herausforderung sind sich langsam bewegende Objekte. Ein neuer Algorithmus wird entwickelt, um die durch langsame Bewegungen verursachte Fehler im Hintergrundbild zu verringern. Zusätzlich wird ein effizienter Algorithmus vorgestellt, durch den der Rechenaufwand signifikant verringert wird.

In Abschnitt 4.4 erfolgt in einem dritten Schritt die eigentliche Segmentierung der Objekte durch einen Vergleich der Bilder der Kameraeinstellung mit dem konstruierten Hintergrundbild. Zur Verringerung von Segmentierungsfehlern wird neben morphologischen Operatoren ein neuer Algorithmus zur Erhöhung der Genauigkeit der Objektgrenzen eingesetzt. Experimentelle Ergebnisse zur Segmentierung und Erzeugung von Hintergrundbildern werden in Abschnitt 4.5 vorgestellt. Damit bei Belichtungsänderungen zwischen den einzelnen Aufnahmen keine Fehler an den Übergängen der Bilder entstehen, wird ein Verfahren zur Verringerung dieser Fehler vorgeschlagen.

4.1 Kamerabewegungen zwischen beliebigen Bildern

Um ein Hintergrundbild aus den Bildern einer Kameraeinstellung zu erzeugen, müssen alle Bilder zunächst passend anhand ihres Bildhintergrundes ausgerichtet werden. Hierbei wird angenommen, dass für zwei aufeinander folgende Bilder einer Kameraeinstellung die Parameter des Kameramodells bekannt sind (vgl. Kapitel 3.4). Ein Bild j wird als Referenzbild festgelegt, um die anderen Bilder an diesem auszurichten. Wird das Bild vor dem Referenzbild mit den Parametern des Kameramodells transformiert, so stimmt – wenn man vom Rauschen und Kompressionsartefakten absieht – der Bildhintergrund beider Bilder überein.

Eine Transformation $\Theta_{i,j}$ zwischen zwei beliebigen Bildern i und j einer Kameraeinstellung

sei durch die acht Parameter des Kameramodells entsprechend Gleichung 3.1 definiert [549]. Aus der Analyse der Kamerabewegung sind zunächst nur die Transformationen $\Theta_{i,i+1}$ zwischen zwei jeweils benachbarten Bildern bekannt, wobei eine unbekannte Transformation mit folgendem Algorithmus aus bekannten Transformationen abgeleitet werden kann: Wählt man beliebige Koordinaten (x, y) im Bild i und transformiert den Punkt mit $\Theta_{i,i+1}$, so wird die Position (x', y') dieses Pixels im Bild $i + 1$ ermittelt. Um die Position des Pixels im Bild $i + 2$ zu erhalten, wird (x', y') mit $\Theta_{i+1,i+2}$ transformiert und ergibt die Position (x'', y'') im Bild $i + 2$. Der Bewegungsvektor von (x, y) nach (x'', y'') entspricht der Verschiebung eines Hintergrundpixels über zwei Bilder. Vier unterschiedliche Punkte werden mit $\Theta_{i,i+1}$ und $\Theta_{i+1,i+2}$ transformiert und ergeben vier Bewegungsvektoren. Durch Einsetzen dieser vier Vektoren in die Gleichung des Kameramodells ist eine eindeutige Berechnung der acht Parameter des Kameramodells von Bild i zu Bild $i + 2$ möglich. Mit dem gleichen Verfahren, d. h. der Auswahl und wiederholten Transformation von vier Punkten, können für beliebige Bilder i und j ($\forall i \leq j$) alle Transformationen $\Theta_{i,j}$ berechnet werden.

Um aus der Transformation $\Theta_{i,j}$ die inverse Transformation $\Theta_{j,i}$ abzuleiten, werden vier Bewegungsvektoren von Bild i nach j durch $\Theta_{i,j}$ bestimmt. Die Richtungen der vier Bewegungsvektoren werden umgedreht, d. h. Startpunkte und Endpunkte werden vertauscht. Die vier Vektoren definieren durch Lösen des Gleichungssystems die Transformation $\Theta_{j,i}$. Aus den bekannten Transformationen $\Theta_{i,i+1}$ von benachbarten Bildern können somit beliebige Transformationen $\Theta_{i,j}$ für alle Bilder einer Kameraeinstellung abgeleitet werden.

4.2 Transformation eines Bildes

Zwei Verfahren zur Transformation eines Bildes werden im Folgenden vorgestellt. Zunächst wird ein beliebiges Bild der Kameraeinstellung als Referenzbild ausgewählt, an dem alle anderen Bilder ausgerichtet werden sollen (engl. *image registration*) [55, 106, 224, 453, 592]. Wird Bild j als Referenzbild festgelegt, so muss jedes Bild i mit $\Theta_{i,j}$ transformiert werden, um einen deckungsgleichen Hintergrund zu erhalten. Zunächst wird die gewünschte Größe des Hintergrundbildes definiert, um anschließend jedem Pixel des Hintergrundbildes einen Pixelwert aus den transformierten Bildern zuzuordnen. Die inverse Transformation $\Theta_{j,i}$ liefert – ausgehend von der Pixelposition (x', y') im Hintergrundbild – die Position (x, y) im ursprünglichen Bild. Die Transformation eines Bildes wird zunächst am Beispiel von Graustufenbildern betrachtet. Jedem Pixel an der Position (x', y') mit $x', y' \in \mathbb{N}$ im transformierten Bild I' wird der Helligkeitwert an der Position (x, y) mit $x, y \in \mathbb{R}$ aus dem ursprünglichen Bild zugewiesen.

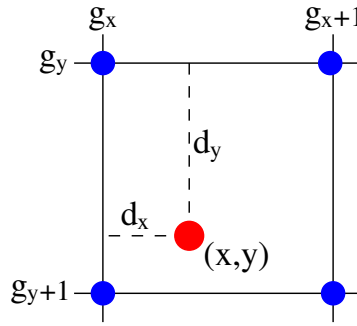


Abbildung 4.1: Lineare Interpolation zur Berechnung der Helligkeit eines Pixels aus benachbarten Pixeln

(x, y) entspricht jedoch nicht genau einer ganzzahligen Pixelposition, sondern wird im Allgemeinen zwischen vier Pixeln liegen. Aus den vier benachbarten Pixeln an den ganzzahligen Pixelpositionen wird der Helligkeitswert des Pixels (x, y) abgeleitet.

Der Helligkeitswert des gesuchten Pixels wird durch lineare Interpolation berechnet. Der ganzzahlige Anteil von x bzw. von y wird mit g_x und g_y bezeichnet, der Rest mit $d_x := x - g_x$ und $d_y := y - g_y$. Die vier ganzzahligen Pixelpositionen um den Punkt (x, y) liegen an den Positionen (g_x, g_y) , $(g_x + 1, g_y)$, $(g_x, g_y + 1)$ und $(g_x + 1, g_y + 1)$ (vgl. Abbildung 4.1). Die Helligkeit I' im transformierten Bild berechnet sich durch Gewichtung der Helligkeitswerte der benachbarten Pixel:

$$I'(x', y') = [(1-d_x) \cdot I(g_x, g_y) + d_x \cdot I(g_x + 1, g_y)] \cdot (1-d_y) + [(1-d_x) \cdot I(g_x, g_y + 1) + d_x \cdot I(g_x + 1, g_y + 1)] \cdot d_y \quad (4.1)$$

Je näher das Pixel (x, y) an einer ganzzahligen Pixelposition liegt, umso geringer ist der Einfluss der anderen Pixel auf die Helligkeit des Pixels. Ein wesentlicher Nachteil der Interpolation ist die Unschärfe, die bis zu einer Verschiebung von 0,5 Pixel zunimmt. Wird ein Bild um 0,5 Pixel horizontal und vertikal verschoben, so entspricht jeder Helligkeitswert des transformierten Bildes dem Durchschnittswert aus jeweils vier Pixeln des Originalbildes, wodurch ein geglättetes Bild entsteht.

Ein zweites Verfahren ermöglicht die Transformation eines Bildes, ohne Unschärfe zu erzeugen. Dabei wird statt des interpolierten Wertes der Helligkeitswert des nächstgelegenen Pixels verwendet. Nachteilig für dieses Verfahren ist eine geringere Genauigkeit der Transformation, da statt einer horizontalen und vertikalen Verschiebung mit Subpixelgenauigkeit nur eine Verschiebung um ganzzahlige Werte möglich ist.

Beide Verfahren, d. h. die Interpolation und die Auswahl des nächstgelegenen Pixels, eignen

sich auch zur *Transformation von Farbbildern*. Hierbei wird jeder Farbkanal einzeln mit dem Kameramodell transformiert. Es hängt im Wesentlichen von der Anwendung ab, ob die Transformation durch Interpolation Vorteile bietet. Bei der bewegungsbasierten Segmentierung von Objekten ist eine möglichst genaue Abbildung der Kamerabewegung von zentraler Bedeutung, so dass die Transformation durch Interpolation erfolgen sollte. Bei der Interpolation werden mehrere Farbwerte miteinander kombiniert, so dass neue Farben im Bild entstehen können. Für hochauflösende Panoramabilder sind Unschärfe und fehlerhafte Farben nicht wünschenswert und die Interpolation somit kein geeignetes Verfahren.

Für Farbbilder bietet sich ein neuer Ansatz an, bei dem die einzelnen Kanäle unterschiedlich transformiert werden. Dazu eignen sich beispielsweise der *HSI*- oder *YUV*-Farbraum [131, 170], in denen Helligkeit und Farbkomponente getrennt kodiert werden. Der *HSI*-Farbraum setzt sich aus dem Farbton *H* (engl. *hue*), der Sättigung *S* (engl. *saturation*) und der Helligkeit *I* (engl. *intensity*) zusammen, die angibt, wie stark eine Farbe mit Weiß gemischt ist. Im *YUV*-Farbraum beschreibt die *Y*-Komponente die Helligkeit und die *UV*-Komponenten die Farben (*Chrominanzwerte*). Die Nachteile der beiden Verfahren können durch eine Kombination der Transformationen deutlich verringert werden. Fehlfarben werden vermieden ohne auf Subpixelgenauigkeit zu verzichten, indem der Helligkeitswert bei der Transformation von Farbbildern interpoliert wird und die Farbwerte durch das nächstgelegene Pixel bestimmt werden.

4.3 Konstruktion von Hintergrundbildern

Nach der Transformation aller Bilder einer Kameraeinstellung unterscheiden sich diese im optimalen Falle nur in den Bereichen mit Objektbewegungen. In einem *Hintergrund*- oder *Panoramabild* sollen Vordergrundobjekte nicht oder höchstens einmal enthalten sein [82, 224, 356, 503]. Alle N Bilder einer Kameraeinstellung sind entsprechend der gewünschten Größe des Hintergrundbildes transformiert worden, so dass für die Beschreibung eines Pixels im Hintergrundbild bis zu N Pixel aus den transformierten Bildern zur Verfügung stehen. Wie in Abbildung 4.2 deutlich wird, verschieben sich durch die Kamerabewegung die Bilder, so dass weniger als N Pixel an einer Bildposition liegen können.

Aus den maximal N Pixeln soll das Pixel gewählt werden, das den Hintergrund möglichst gut beschreibt. Unter der Annahme, dass mindestens die Hälfte der Pixel den Bildhintergrund beschreibt, liefert der *Median* dieser N Helligkeitswerte eine gute Heuristik für ein Hintergrundpixel. Abbildung 4.2 verdeutlicht schematisch die Konstruktion des Bildhintergrundes.

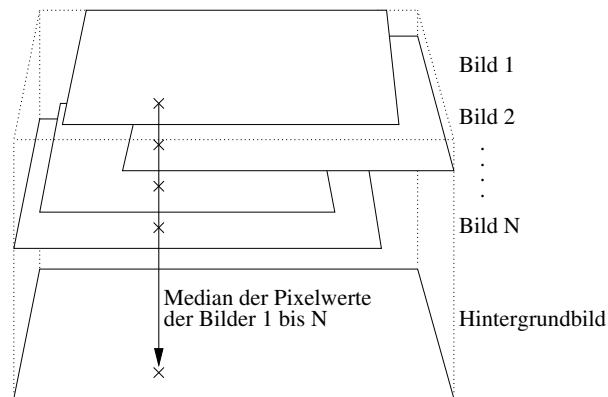


Abbildung 4.2: Die Helligkeit eines Pixels im Bildhintergrund wird durch den Median der Pixelwerte an einer Bildposition der transformierten Bilder 1 . . . N bestimmt.

Nach der Transformation aller Bilder der Kameraeinstellung wird der Median an jeder Pixelposition berechnet und definiert das Hintergrundbild.

In vielen Videosequenzen bewegt sich ein Objekt so langsam durch das Bild, dass einzelne Objektpixel mehr als die Hälfte der Zeit an einer Pixelposition verweilen. Der Median wählt dann für den Bildhintergrund Objektpixel aus, so dass fehlerhafte Bereiche in den Hintergrundbildern entstehen. In Abbildung 4.3 bewegt sich eine Person so langsam durch das Bild, dass ein Fuß, der für einen längeren Zeitraum an einer Bildposition verweilt, Teil des Hintergrundbildes wird.

Ein verbesserter neuer Algorithmus wird im Folgenden vorgeschlagen, um diese Artefakte zu vermeiden. Statt des Medians werden zunächst Differenzbilder durch einen direkten Vergleich zweier benachbarter und durch die Transformation entsprechend ausgerichteter Bilder berechnet. Bei der Bewegung eines Objektes treten deutliche Bildunterschiede in mindestens zwei Regionen auf. So ist nach der Bewegung des Objektes ein Teil des Hintergrundes verdeckt, und ein Teil des zuvor verdeckten Hintergrundes wird sichtbar. In Abbildung 4.4, in der eine Person zu zwei unterschiedlichen Zeitpunkten innerhalb einer Kameraeinstellung abgebildet ist, sind die Regionen mit signifikanten Pixeldifferenzen gelb markiert. In der Nähe der Objektgrenzen liegen viele Pixel mit starken Pixeldifferenzen, und nur vereinzelt treten hohe Differenzen durch Rauschen in anderen Bildbereichen auf.

Aus den Bildbereichen mit den starken Pixeldifferenzen kann die Position und Größe des Objektes geschätzt werden. Dazu wird zunächst angenommen, dass sich genau ein Objekt im Bild bewegt. Das Differenzbild wird in ein Binärbild $D(x, y) \in \{0, 1\}$ umgewandelt, in dem Pixel mit einem hohen absoluten Differenzwert durch eine 1 repräsentiert sind. Der Schwer-



Abbildung 4.3: Fehlerhaftes Hintergrundbild bei langsamer Objektbewegung

punkt (S_x, S_y) der markierten Differenzpixel liefert eine gute und sehr effizient zu berechnende Schätzung für die Position des Objektes:

$$S_x = \frac{1}{\sum_{x,y} D(x,y)} \sum_{x,y} x \cdot D(x,y) \quad (4.2)$$

$$S_y = \frac{1}{\sum_{x,y} D(x,y)} \sum_{x,y} y \cdot D(x,y) \quad (4.3)$$

Zentriert um den Schwerpunkt wird ein Rechteck der Breite $R_x = n \cdot \sigma_x$ und Höhe $R_y = n \cdot \sigma_y$ gelegt. σ_x und σ_y bezeichnen die Varianzen der Pixelpositionen der x- bzw. y-Koordinate der markierten Differenzpixel. Der konstante Faktor n skaliert die Größe des Rechtecks, wobei gute Ergebnisse mit Werten im Intervall $[2, 4]$ erzielt werden. Die Pixel innerhalb des durch Breite, Höhe und Zentrum definierten Rechtecks sind mit hoher Wahrscheinlichkeit Objektpixel. Abbildung 4.4 gibt für die markierten Differenzpixel die geschätzte Position des Objektes an. Die Festlegung des rechteckigen Bereiches bietet zudem den Vorteil, dass auch Pixel innerhalb des Objektes erfasst werden, die sich zwischen zwei benachbarten Bildern nicht verändern. Die Pixel des rechteckigen Bereiches sollen keinen bzw. nur einen geringeren Einfluss auf das konstruierte Hintergrundbild haben und können als Hintergrundpixel ausgeschlossen oder während der Berechnung geringer gewichtet werden. Bei einer Gewichtung wird ein Pixel



Abbildung 4.4: Die signifikanten Differenzen zwischen zwei transformierten Bildern wurden markiert. Das Rechteck wird durch den Schwerpunkt und die Varianz der Positionen der Differenzpixel definiert.

innerhalb des Rechtecks einfach und alle Pixel außerhalb mehrfach für den Median berücksichtigt. So kann sichergestellt werden, dass für jede Bildposition mindestens ein Pixel zur Verfügung steht und ein Hintergrundbild ohne Lücken konstruiert wird.

Eine weitere Verbesserung des Verfahrens ist möglich, indem mehrere unterschiedlich große Rechtecke um den Schwerpunkt berücksichtigt werden. Dabei wird die Anzahl der zur Berechnung des Medians verwendeten Pixel anhand der durch den Faktor n definierten Größe des umgebenden Rechtecks bestimmt. Pixel nahe am Schwerpunkt liegen im kleinsten durch $n = 1$ definierten Rechteck und werden bei der Berechnung des Medians nur einfach berücksichtigt. Die Gewichtung steigt mit zunehmender Entfernung beziehungsweise steigendem n .

Bei mehreren Objekten im Bild ist eine Erweiterung des vorgestellten Verfahrens erforderlich. In einem ersten Schritt wird der Schwerpunkt der Differenzpixel berechnet. Falls die Varianzen der Pixelpositionen der Differenzpixel einen Schwellwert überschreiten, wird angenommen, dass mindestens zwei Objekte im Bild enthalten sind. In diesem Fall werden die Differenzpixel mit Hilfe des *K-Means-Algorithmus* in zwei Gruppen eingeteilt und jede Gruppe erneut

analysiert und deren Schwerpunkt bestimmt. Die Unterteilung wird iterativ fortgesetzt, bis die Varianzen den Schwellwert nicht mehr übersteigen.

Der *Rechenaufwand zur Bestimmung des Medians* ist sehr hoch, da er für jedes Pixel des Hintergrundbildes berechnet wird und einzelne Pixel – sofern diese in größerer Entfernung zum Schwerpunkt liegen – mehrfach berücksichtigt werden. In einer Liste mit aufsteigend sortierten Pixelwerten entspricht der Median dem mittleren Wert. Durch die Sortierung der Liste liegt die Komplexität des Algorithmus bei $O(n \log n)$ [26, 119].

Der folgende deutlich effizientere Algorithmus reduziert die Komplexität auf $O(n)$. Statt eine Liste zu sortieren und den mittleren Wert auszuwählen, wird aus den Pixelwerten an einer Bildposition ein Histogramm erzeugt. Der jeweilige Histogrammwert wird um eins erhöht, falls die Position innerhalb des durch $n = 1$ definierten Rechtecks liegt, mit zunehmender Entfernung abhängig vom Faktor n bis zu einem Wert von fünf. Zur Ermittlung des Medians wird das Histogramm aufsteigend durchlaufen und die Anzahl der Pixel summiert. Der Median entspricht dem Indexeintrag, bei dem die Summe die Hälfte aller Indexwerte des Histogramms überschreitet. Insbesondere in längeren Kameraeinstellungen mit geringer Kamerabewegung beschleunigt die Verwendung von Histogrammen die Rechenzeit des Medians signifikant.

Das Verfahren zur Erzeugung von Hintergrundbildern eignet sich nicht nur für Videos, sondern kann auch zur Erzeugung von Panoramabildern aus Digitalfotografien verwendet werden. Das Kameramodell basierend auf acht Parametern ermöglicht im Gegensatz zum zylindrischen oder sphärischen Kameramodell die korrekte Berechnung der Panoramabilder auch bei perspektivischen Verzerrungen oder Rotationen der Kamera entlang der Blickrichtung, die bei Aufnahmen ohne Verwendung eines Stativs häufig zu beobachten sind. Ein weiteres Einsatzgebiet sind *Background-Sprites* in MPEG-4 komprimierten Videos [226], in denen zur Reduktion der Bitrate das Hintergrundbild getrennt kodiert und übertragen wird [225, 502, 530].

4.4 Segmentierung von Objekten

Durch die Anwendung des Medianfilters sind im konstruierten Hintergrundbild die Objekte des Vordergrundes nicht mehr enthalten. Die Segmentierung eines Objektes erfolgt durch den Vergleich des transformierten Bildes mit dem Hintergrundbild. Unter der Annahme, dass sich das Objekt deutlich vom Hintergrund unterscheidet, kann dessen Position und Form exakt bestimmt und als Binärbild gespeichert werden. Um die Position und Form im ursprünglichen Bild der Kameraeinstellung zu ermitteln, wird das Binärbild mit Hilfe der inversen Transformation auf das ursprüngliche Bild transformiert.

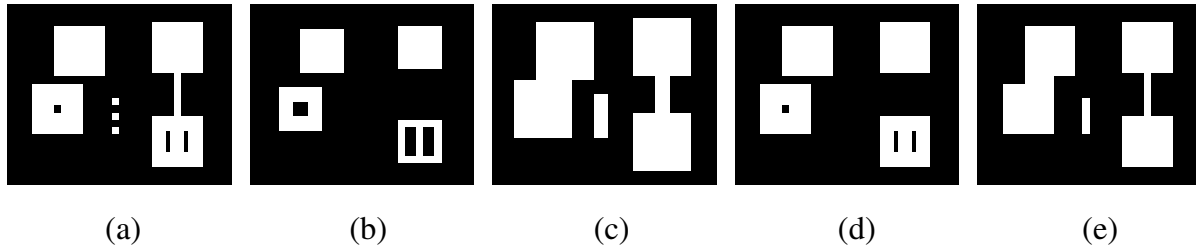


Abbildung 4.5: Morphologische Operatoren: Originalbild (a), Erosion (b), Dilatation (c), Opening (d) und Closing (e).

Rauschen, Kompressionsartefakte und geringe Fehler bei der Berechnung des Kameramodells können deutlich sichtbare Fehler im Differenzbild verursachen. Um diese Fehler auszugleichen, wird das transformierte Differenzbild durch *morphologische Operatoren* geglättet [53, 120, 537]. Die beiden Operatoren *Dilatation* und *Erosion* sind für ein Strukturelement B und ein Graustufenbild I definiert als:

$$\text{Dilatation : } D_B(I(x)) = \max \{I(x+r) \mid r \in B\}, \quad (4.4)$$

$$\text{Erosion : } E_B(I(x)) = \min \{I(x+r) \mid r \in B\}. \quad (4.5)$$

Üblicherweise werden als *Strukturelemente* Kreise, Ellipsen oder Rechtecke verwendet. Am Beispiel des Binärbildes in Abbildung 4.5 (a) werden die Auswirkungen der Operatoren verdeutlicht. Die Erosion trägt Ränder von Objekten ab, wogegen die Dilatation Objekte vergrößert und Lücken zwischen Objekten schließt.

Die beiden abgeleiteten Operatoren *Opening* und *Closing*¹ [476] kombinieren Dilatation und Erosion:

$$\text{Opening : } O_B(x) = D_B [E_B (I(x))], \quad (4.6)$$

$$\text{Closing : } C_B(x) = E_B [D_B (I(x))]. \quad (4.7)$$

Während beim *Opening*-Operator zunächst eine Erosion mit anschließender Dilatation erfolgt, ist die Reihenfolge beim *Closing*-Operator umgekehrt. Durch die Glättung des *Opening*-Operators werden kleine und schmale Objektregionen entfernt, größere Regionen bleiben jedoch weitgehend unverändert erhalten. Der *Closing*-Operator füllt Löcher und schließt Lücken zwischen benachbarten Regionen [189].

¹Die englischen Begriffe für die morphologischen Operatoren *Opening* (öffnen) und *Closing* (schließen) haben sich im Deutschen als Fachbegriffe durchgesetzt.



Abbildung 4.6: *Ergebnisse der Segmentierung: Differenzbild aus transformiertem Bild und Hintergrundbild (a), Differenzbild nach Anwendung morphologischer Operatoren und Auswahl des größten Objektes (b) und Optimierung der Ränder der segmentierten Person (c).*

Die Anwendung beider abgeleiteter Operatoren auf ein Graustufenbild kombiniert die Vorteile der Verfahren. Zunächst entfernt der *Opening*-Operator kleine Regionen im Differenzbild wie z. B. einzelne durch Rauschen veränderte Pixel. Der *Closing*-Operator schließt anschließend Lücken innerhalb eines Objektes und zwischen angrenzenden Objektregionen, so dass die Qualität der Differenzbilder signifikant verbessert wird.

In einem letzten Schritt wird der äußere Rand des Objektes analysiert, um Segmentierungsfehler zu verringern. Morphologische Operatoren glätten die Ränder der segmentierten Objekte. Um die Auswirkung der Glättung zu reduzieren, werden starke Kanten in den Randbereichen eines Objektes gesucht, wobei die Art und Größe des Strukturelements des morphologischen Operators eine Abschätzung über die Änderung der Kontur ermöglicht. Im Randbereich der äußeren Kontur werden starke Kanten markiert. Falls keine starke Kante in der Nähe eines Konturpixels gefunden wird, bleibt die ursprüngliche Objektgrenze unverändert, ansonsten wird der Rand entsprechend vergrößert oder verkleinert.

Einen Überblick über die Ergebnisse der einzelnen Segmentierungsschritte gibt Abbildung 4.6. Die Analyse der Kanten im Randbereich des Objektes liefert insbesondere in Regionen mit stark ausgeprägten Kanten deutlich genauere Segmentierungsergebnisse bei den analysierten Videos.

Dargestelltes Objekt	Anzahl Bilder	Kameraoperation	Faktor Rechenzeit für Median
<i>Tennispieler I</i>	100	horizontaler Schwenk	13,9
<i>Tennispieler II</i>	80	Zoom-In, vertikaler Schwenk	10,4
<i>Person I</i>	300	verwackelte Kamera	30,7
<i>Person II</i>	65	horizontaler Schwenk	11,6
<i>Person III</i>	30	horizontaler Schwenk	5,4
<i>PKW an Ampel</i>	60	diagonaler Schwenk	15,7
<i>Lieferwagen</i>	105	Zoom-In	17,4
<i>Rennwagen</i>	45	horizontaler Schwenk	7,6
<i>Katze</i>	50	diagonaler Schwenk	10,3
<i>Schiff</i>	300	horizontaler Schwenk	26,2

Tabelle 4.1: Testsequenzen zur automatischen Objektsegmentierung. Die Werte der letzten Spalte geben an, um welchen Faktor die Berechnung des Medians durch die Verbesserung des Verfahrens beschleunigt wird.

4.5 Experimentelle Ergebnisse

Die Qualität der Segmentierung hängt im Wesentlichen von der Genauigkeit des berechneten Kameramodells und des daraus abgeleiteten Hintergrundbildes ab. Zehn kurze Kameraeinstellungen mit einer Länge zwischen 30 und 300 Bildern werden analysiert. In jeder Kameraeinstellung sind Objekt- und Kamerabewegungen enthalten. Tabelle 4.1 gibt einen Überblick über die analysierten Bildsequenzen und verdeutlicht, um welchen Faktor die Berechnung des Medians bei der Verwendung des effizienteren Verfahrens für die einzelnen Bildsequenzen beschleunigt wird.

In allen Sequenzen ist die Segmentierung des jeweils abgebildeten Objektes möglich. Das Objekt nimmt nur einen kleinen Teil der Bildfläche ein, und markante Strukturen im Bildhintergrund, durch die viele Ecken eindeutig festgelegt werden können, ermöglichen eine gute Schätzung der Parameter des Kameramodells und die korrekte Berechnung der Hintergrundbilder.

In mehreren Sequenzen werden die Objekte einzelner Bilder fehlerhaft segmentiert. Diese Fehler sind insbesondere in den ersten oder letzten Bildern einer Kameraeinstellung zu beobachten, falls sich das Objekt in geringer Entfernung zur Kamera befindet und einen großen Teil des Bildes ausfüllt. Dadurch entstehen fehlerhafte Transformationen, die automatisch erkannt werden, so dass die entsprechenden Bilder bei der Berechnung des Hintergrundbildes und der Segmentierung unberücksichtigt bleiben.

Ein weiterer mehrfach zu beobachtender Fehler entsteht durch den Schatten eines Objektes. In

den betroffenen Bildbereichen ändert sich die Helligkeit der Hintergrundpixel, so dass Teile des Schattens gemeinsam mit dem Objekt segmentiert werden. In Abbildung 4.7 treten in den Sequenzen *Rennwagen*, *Person I* und *Katze* Segmentierungsfehler in den schattigen Regionen auf.

Vereinzelte ähneln sich die Helligkeits- bzw. Farbwerte von Objekt und Hintergrund, so dass im Differenzbild keine Unterschiede erkennbar sind und das Objekt nicht vollständig segmentiert wird. Kleine fehlerhafte Regionen werden durch die Glättung mit den morphologischen Operatoren entfernt. Rauschen und geringe Veränderungen im Bildhintergrund erzeugen ebenfalls Fehler bei der Segmentierung. Eine Mindestgröße für Objekte bzw. die Auswahl des größten Objektes im Bild verhindert diese Fehler, die insbesondere bei Filmaufnahmen in geringer Qualität auftreten. Beispiele für automatisch segmentierte Objekte der einzelnen Testsequenzen sind in Abbildung 4.7 dargestellt. Anhand der markierten Objektregionen wird deutlich, dass Segmentierungsfehler in den Randbereichen eines Objektes und insbesondere in Bereichen mit Schatten auftreten können.

Neben der Segmentierung wird analysiert, wie gut sich das Verfahren zur Erzeugung von Panoramabildern aus Videosequenzen und Einzelbildern eignet. Videoaufnahmen, die speziell für Panoramabilder erzeugt werden, enthalten nur selten große Objekte im Bildvordergrund, wodurch in den analysierten Testsequenzen deutlich weniger Fehler zu beobachten sind. Ein Problem bei der Verwendung von Einzelbildern sind insbesondere die Übergänge an den Bildgrenzen der transformierten Bilder, die deutlich sichtbare Artefakte im Panoramabild hinterlassen können. Ändern sich die Lichtverhältnisse bzw. die Belichtung zwischen den Aufnahmen, so können sich die Farbwerte an einer Pixelposition in den transformierten Bildern deutlich unterscheiden. Ein fließender und gleichmäßiger Übergang zwischen den aneinander grenzenden Bildbereichen ist durch die Berechnung des Medians nicht immer möglich, da bei Panoramabildern aus Digitalfotografien häufig nur zwei oder drei Bilder überlappen. Zur Erkennung und Reduzierung möglicher Bildfehler werden in den transformierten Bildern zunächst deutliche Pixeldifferenzen in den überlappenden Bereichen identifiziert. Bei wesentlichen Unterschieden der Pixelwerte werden diese Bereiche senkrecht zur erwarteten Kante geglättet. Abbildung 4.7 zeigt zwei aus Einzelbildern automatisch erzeugte Panoramabilder, in denen die Übergänge zwischen den Bildern automatisch geglättet wurden.

Am Beispiel von Abbildung 4.8 wird deutlich, dass die automatische Segmentierung von Objekten in Videosequenzen neue Möglichkeiten zur nachträglichen Änderung bestehender Filme eröffnet. Beispielhaft wird ein Rennwagen eines historischen Schwarz-Weiß-Videos in eine aktuelle Videosequenz eingefügt, wobei die Kamerabewegung des neuen Videos der Kamera-

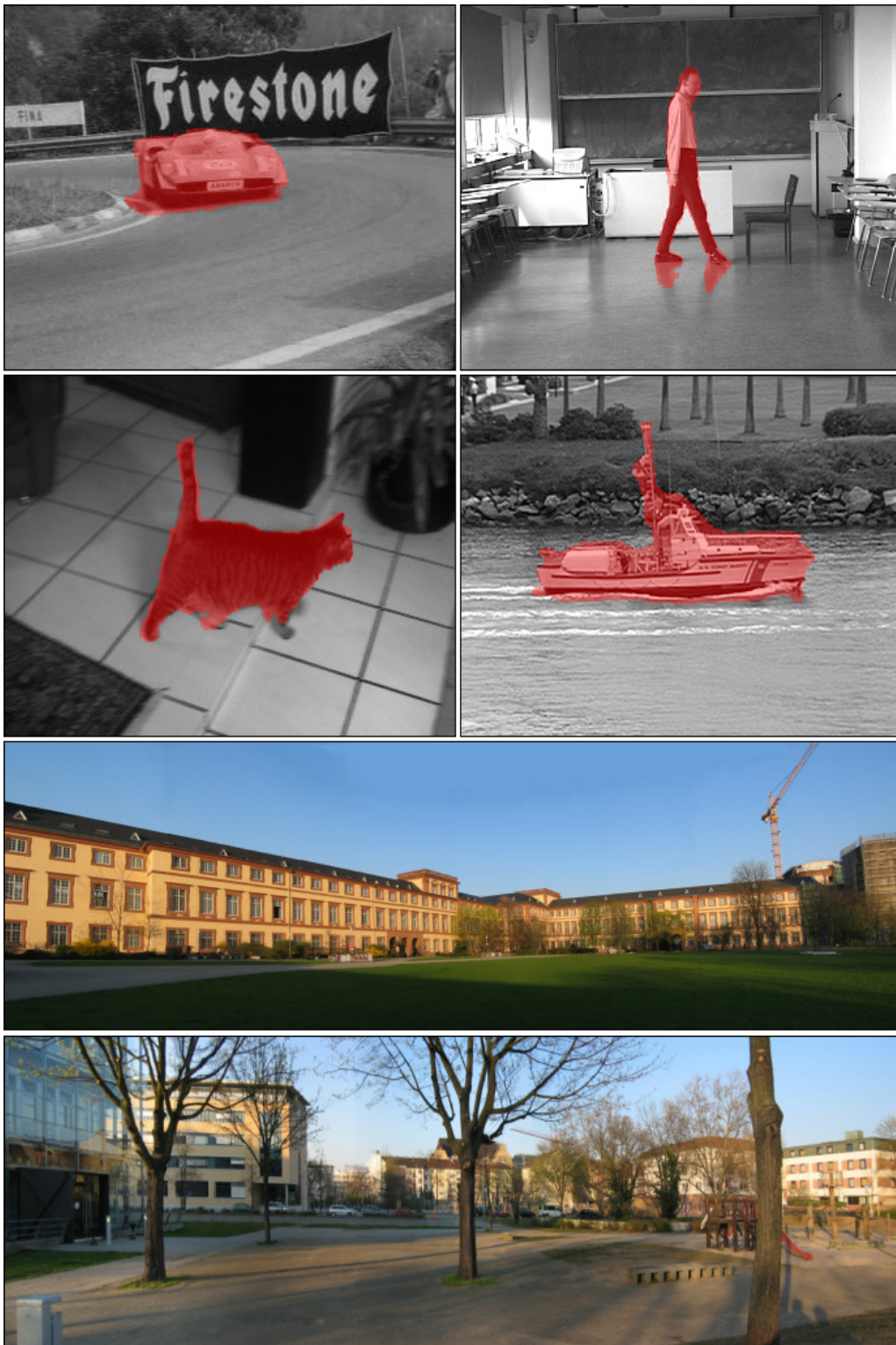


Abbildung 4.7: Oben: Beispiele für automatisch segmentierte Objekte der Testsequenzen "Rennwagen", "Person I", "Katze" und "Schiff". Unten: Automatisch erzeugte Panoramabilder.

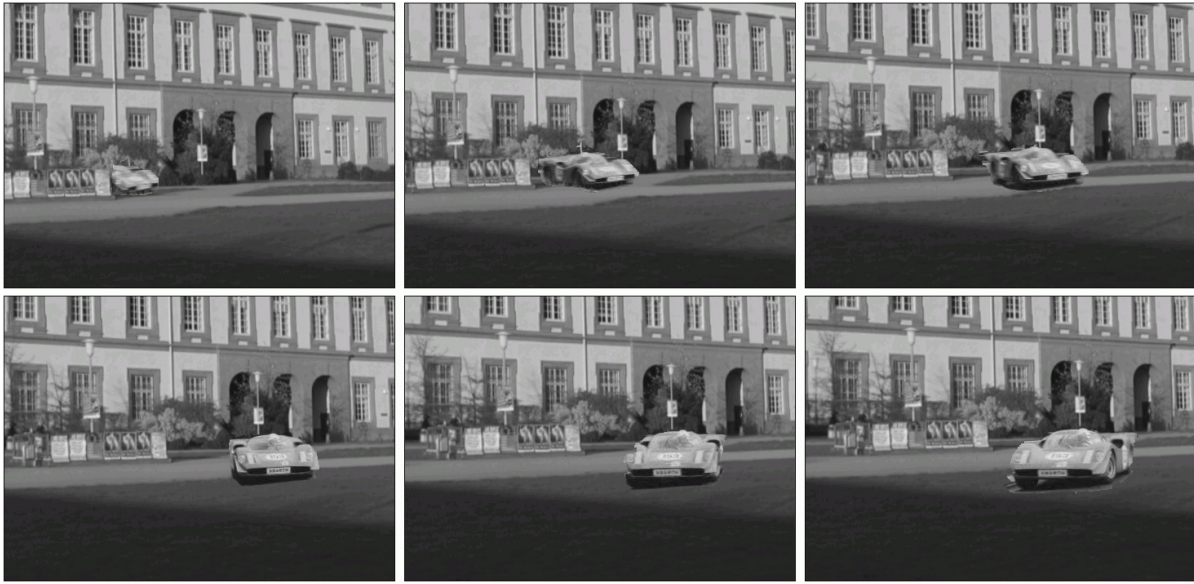


Abbildung 4.8: *Nachträgliches Einfügen von Objekten in Videosequenzen: Ein Rennwagen aus einer historischen Dokumentation wird nach Festlegung der Position und des Skalierungsfaktors automatisch in eine aktuelle Videosequenz eingefügt. Die Vordergrundobjekte des aktuellen Videos werden dabei entfernt.*

bewegung des historischen Videos entspricht. Da keine Farbinformationen über den Rennwagen vorliegen, wird das neue Video als Schwarz-Weiß-Videos gespeichert. Zu Beginn der Kameraeinstellung fährt der Rennwagen in das Bild, so dass die manuelle Auswahl einer geeigneten Startposition des Rennwagens von besonderer Bedeutung ist. Im letzten Bild sind Segmentierungsfehler sichtbar, die insbesondere durch den Schatten des Autos verursacht werden.

Durch die automatische Segmentierung ist es ohne größeren manuellen Aufwand möglich, Objekte aus einer Sequenz auszuschneiden und in ein zweites Video einzufügen. Es muss lediglich darauf geachtet werden, dass die Position und Größe des Objektes zum Inhalt des zweiten Filmes passt und die Lichtverhältnisse beider Kameraeinstellungen einander entsprechen. In Abbildung 4.8 wurden die Anfangsposition und der Skalierungsfaktor des Rennwagens manuell für das erste Bild der Videosequenz festgelegt. An den unterschiedlichen Richtungen des Schattens wird deutlich, dass die Lichtverhältnisse in diesem Beispiel nicht berücksichtigt werden.

4.6 Zusammenfassung

In diesem Kapitel wurde ein Verfahren zur Objektsegmentierung durch Bewegungsanalyse vorgestellt. Ein Hintergrundbild wurde erzeugt, indem die Bilder einer Kameraeinstellung entsprechend ausgerichtet und Vordergrundobjekte durch Berechnung des Medians entfernt wurden. Ein neuer Algorithmus wurde entwickelt, der insbesondere bei langsamen Objektbewegungen Fehler im Hintergrundbild deutlich reduziert und dennoch eine effiziente Berechnung des Medians ermöglicht. Die Segmentierung eines Objektes erfolgte durch den Vergleich der Bilder der Kameraeinstellung mit dem Hintergrundbild. Anschließend wurde ein neues Verfahren zur Verringerung von Segmentierungsfehlern durch Analyse der Kanten im Bereich der äußeren Kontur des Objektes vorgestellt. Experimentelle Ergebnisse zur Objektsegmentierung, zur Erzeugung von Panoramabildern und zum nachträglichen Einfügen von Objekten in andere Videosequenzen ergänzen das Kapitel.

KAPITEL 5

Klassifikation von Objekten

Die Erkennung eines Objektes ist ein wichtiger Schritt in der automatischen Analyse von Videos. Objekte liefern semantische Informationen, die insbesondere zur Indexierung von Videodatenbanken und für eine Suche nach speziellen Videosequenzen herangezogen werden können [56, 98, 369, 395, 516]. Aufgrund der semantischen Bedeutung von Objekten bietet es sich auch an, computergenerierte Zusammenfassungen von Videos auf Grundlage der erkannten Objekte zu erzeugen [256, 281].

Ein Mensch kann ein und dasselbe Objekt auf verschiedenen Ebenen erkennen und beschreiben [64, 498]. Die höchste Ebene, die ein Mensch auch am schnellsten wahrnimmt, ist die Ebene der *Objektklasse*, in der mehrere gleichartige Objekte in einer übergeordneten Kategorie zusammengefasst werden. Beim Betrachten eines Bildes fällt einem Menschen der Name der Objektklasse spontan ein, wie z. B. die Objektklasse *Vogel* oder *Mensch*. Erst bei genauerer Betrachtung können *spezielle Eigenschaften* [496] des Objektes anhand von Textur- und Farbinformationen erkannt werden [105, 326]. Hierzu zählen beispielsweise Tierarten wie eine *Amsel*. Die *individuelle Objektbezeichnung* identifiziert ein bekanntes und individuell benanntes Objekt ("die Katze meines Nachbarn") und erfordert detaillierte Kenntnisse über das Aussehen des Objektes.

Ziel der automatischen Klassifikation soll im Folgenden die *Erkennung der Objektklasse* sein. Ein höherer Detaillierungsgrad würde speziell angepasste Datenbanken erfordern, und um die Größe der Datenbank zu beschränken, müsste die Erkennung auf wenige Objekte eingeschränkt werden.

Innerhalb der Wahrnehmungspsychologie wurden eine Reihe von Theorien über die Art der Repräsentation von Objekten im menschlichen Gehirn entwickelt [36, 347, 520]. Obwohl sich bisher keine einheitliche Theorie durchgesetzt hat, scheinen dreidimensionale Objekte als zweidimensionale Ansichten abgebildet zu werden [495]. Die Drehung eines Objektes zur Kamera hat starken Einfluss, ob und wie schnell ein Mensch ein Objekt erkennt. Einfach zu erkennende zweidimensionale Projektionen eines dreidimensionalen Objektes werden als *kanonische Sichten* (engl. *canonical view*) bezeichnet [404]. Besonders gut geeignet sind Ansichten im Profil oder leicht erhöhte Ansichten von schräg vorne [64]. Weiterhin sind vertraute Perspektiven für die Erkennung besonders vorteilhaft, also Perspektiven, aus denen ein Objekt üblicherweise betrachtet oder im Fall von Gebrauchsgegenständen verwendet wird [40]. Insbesondere in Abschnitt 5.9.1 bei der Auswahl von Objekten für die Referenzdatenbank sollten kanonische Sichten mit Vorrang berücksichtigt werden.

Verfahren zur Beschreibung und Erkennung von Objekten wurden auch bei der Standardisierung von MPEG-7 berücksichtigt [124]. Objekte liefern Informationen über die Inhalte von multimedialen Daten und können dadurch die Suche, den Zugriff und die Adaption von Videos unterstützen. Für die Objekterkennung sind insbesondere die *visuellen Deskriptoren* (engl. *visual descriptor*) wie beispielsweise Farben, Texturen, Bewegungen oder Objektkonturen aus dem dritten Teil des MPEG-7 Standards relevant [41, 227].

Zur Erkennung eines Objektes reicht es für einen Menschen häufig aus, die Kontur und die Änderung der Kontur bei der Bewegung des Objektes zu betrachten [497]. Abbildung 5.1 verdeutlicht, dass die automatisch segmentierten Konturen einer Person durch die Änderungen im Zeitablauf trotz fehlerhafter Segmentierung von einem Menschen leicht erkannt werden können. Dagegen enthalten Farben oder Texturen häufig nicht ausreichend Informationen zur Charakterisierung unterschiedlicher Objekte.

Schon im Jahr 1978 stellte Parlidis eine Übersicht über Verfahren zur Klassifikation von Konturen vor [413]. Trotz der großen Anzahl verfügbarer Ansätze [17, 93, 108, 328, 454, 471] ist das Problem der zuverlässigen Erkennung von Konturen bis heute noch nicht zufriedenstellend gelöst [295, 448, 504, 575]. Um möglichst gute Konturdeskriptoren zu identifizieren, wurden mehrere Verfahren – insbesondere Verfahren basierend auf Wavelets [234], Polygonen [298], Fourriedeskriptoren [582], Eigenvektoren [258] und Skalenraumbildern [366] – bei der Entwicklung des MPEG-7 Standards vorgeschlagen, analysiert und umfangreichen Tests unterworfen [364]. Das Ergebniss der umfangreichen Analysen lässt sich folgendermaßen zusammenfassen [364]: Der Skalenraumansatz führt zu signifikant besseren Klassifikationsergebnissen im Vergleich zu allen anderen analysierten Verfahren. Zudem wird ein Objekt mit

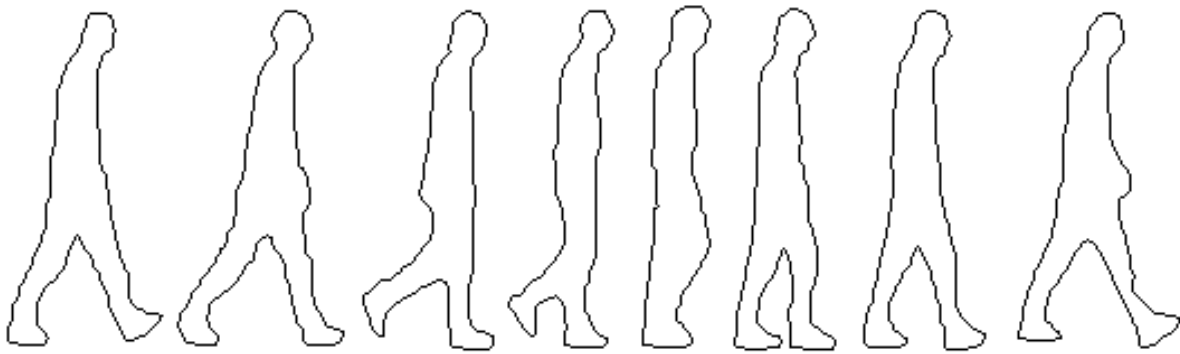


Abbildung 5.1: *Kontur einer Person im Zeitablauf*

einem Datensatz von nur 14 Byte deutlich kompakter beschrieben. Dieser sehr erfolgsversprechende skalenraumbasierte Ansatz wurde im MPEG-7-Standard zur Beschreibung der äußeren Kontur von Objekten ausgewählt.

Ein wesentlicher Vorteil des Skalenraumansatzes besteht darin, dass dieser die menschliche Wahrnehmung bei der Beurteilung der Ähnlichkeiten zweier Konturen sehr gut abbildet [364]. Ein ganz wesentliches Kriterium für einen Menschen bei der Erkennung von Konturen ist die Einteilung in konkave und konvexe Bereiche, die in Skalenraumabbildungen detailliert abgebildet werden. Weiterhin liefert das Verfahren gute Ergebnisse bei teilweiser Verdeckung eines Objektes und ist sehr robust gegenüber Verformungen von Objekten, was beispielsweise für die Erkennung von Personen in Videos besonders wichtig ist. Zudem ist das Verfahren invariant gegenüber Rotationen und Spiegelungen und sehr robust bei Rauschen, bei perspektivischen Verzerrungen und gegenüber der Anzahl und der Auswahl von Konturpixeln. Da zudem noch eine effiziente Berechnung der Merkmalswerte möglich ist, bildet das Verfahren die Grundlage für die Objekterkennung in diesem Kapitel. Trotz der Vorteile enthält der Skalenraumansatz auch deutliche Schwächen, auf die detailliert in diesem Kapitel eingegangen wird. Anschließend stellen wir die von uns neu entwickelten Verfahren vor, durch die eine wesentlich zuverlässigere Objekterkennung möglich wird.

Bei der Klassifikation eines Objektes mit Hilfe des Skalenraumansatzes werden Merkmale, die aus der äußeren Kontur des Objektes abgeleitet werden, miteinander verglichen. Nach der Parametrisierung der Kontur in Abschnitt 5.1 werden globale Konturdeskriptoren vorgestellt, die eine erste Abschätzung der Ähnlichkeit zweier Objekte ermöglichen. Die Erkennung eines Objektes erfolgt durch einen Vergleich von Skalenraumabbildungen, auf die in den Abschnitten 5.3 bis 5.5 eingegangen wird. Zwei wesentliche Probleme bleiben bei dem ursprünglichen Ska-

lenraumvergleich unberücksichtigt: Zum Einen können unterschiedliche konkave Regionen zu identischen Merkmalswerten in Skalenraumabbildungen führen. Wir schlagen in Abschnitt 5.6 ein neues Verfahren zur *Reduktion dieser Mehrdeutigkeiten* vor. Ein zweites wesentliches Problem bei der Objektklassifikation mit Skalenraumabbildungen ist darauf zurückzuführen, dass konvexe Objektregionen nicht berücksichtigt werden und so wichtige Informationen einer Kontur verloren gehen. In Abschnitt 5.7 führen wir das neue Konzept der *transformierten Kontur* ein, durch die erst eine Charakterisierung konvexer Objektregionen möglich wird. Anschließend wird in Abschnitt 5.8 der Begriff der Distanz zwischen Objekt und Objektklasse erläutert und ein neues Verfahren zur Aggregation der Ergebnisse für Videosequenzen vorgestellt. Im Rahmen der experimentellen Ergebnisse werden neben der Datenbank und den Testsequenzen typische Fehlerquellen bei der Objektklassifikation analysiert. Zusätzlich werden Ergebnisse zur Objekterkennung in historischen Videos vorgestellt, die im Rahmen des Projektes *European Chronicles Online* gesammelt wurden.

5.1 Parametrisierung der Kontur

Die äußere Kontur eines Objektes soll durch N Wertepaare $(x(i), y(i))$ mit $i = 0 \dots N - 1$ beschrieben werden. Zur Parametrisierung der Kontur wird ein beliebiger Punkt auf der Kontur als Startposition $(x(0), y(0))$ gewählt. Die Kontur wird im Uhrzeigersinn abgelaufen, und die Positionen aller Konturpixel werden in einer Liste H mit N_H Elementen gespeichert.

Größenunterschiede eines Objektes im Bild, die aus der Einstellung und Entfernung der Kamera resultieren, erzeugen Konturen unterschiedlicher Länge. Zur Klassifikation einer Kontur werden genau N Konturpixel benötigt, d. h. es müssen ggf. Pixel aus der Liste der abgetasteten Konturpixel entfernt bzw. neue hinzugefügt werden. Ist das segmentierte Objekt sehr klein ($N_H < N$), so wird die Anzahl der segmentierten Konturpixel durch Interpolation aus benachbarten Pixeln künstlich erhöht. Bei großen Objekten werden Konturpixel in gleichmäßigen Abständen aus der Liste H gelöscht.

Nach der Normalisierung wird jede Kontur durch genau N Wertepaare beschrieben. Die Konturpixel $(x(0), y(0))$ und $(x(N - 1), y(N - 1))$ liegen benachbart, wobei das Startpixel ein beliebiges Pixel der Kontur ist. Die in den folgenden Abschnitten vorgestellten Verfahren zum Vergleich von Konturen sind rotationsinvariant, so dass die Wahl des Startpunktes keine Auswirkung auf die Klassifikationsergebnisse hat.

5.2 Globale geometrische Konturdeskriptoren

Globale Konturdeskriptoren betrachten die Kontur als Ganzes und beschreiben sie mit einem aggregierten Wert. Sie eignen sich nur für eine grobe Abschätzung der Ähnlichkeit zweier Konturen. Die Aussagekraft dieser Deskriptoren darf nicht zu hoch eingeschätzt werden, da wesentliche Informationen über die ursprüngliche Objektform verloren gehen. Ein Vorteil liegt in ihrer schnellen Berechenbarkeit, so dass signifikante Unterschiede zwischen Konturen schnell erkannt werden können [123, 385]. Betrachtet werden im Folgenden die beiden Maße *Kompaktheit* und *Exzentrizität*.

Die Kompaktheit (engl. *compactness*) eines Objektes beschreibt die Ähnlichkeit einer Kontur mit einem Kreis [470]. Im segmentierten Objekt i wird die Kompaktheit c_i durch die Länge der Kontur U und der Fläche F des Objektes bestimmt. Unterschiede zwischen zwei Konturen i und j in Bezug auf die Kompaktheit α_c werden auf das Intervall $[0, 1]$ normiert:

$$c_i = \frac{U^2}{4 \cdot \pi \cdot F} \quad (5.1)$$

$$\alpha_c(i, j) = \frac{|c_i - c_j|}{\max(c_i, c_j)} \quad (5.2)$$

Die Kompaktheit ist invariant gegenüber geometrischen Transformationen wie Rotation oder Skalierung und kann sehr effizient aus den segmentierten Binärbildern berechnet werden. Der Wert für die Kompaktheit wird bei einem Kreis minimal.

Das nach Brown benannte Maß für die Exzentrizität (engl. *eccentricity*) beschreibt das Verhältnis der Längen der Hauptachsen bezogen auf die zentralen Momente der Konturpixel [22]. Die Exzentrizität e_i der Kontur i wird durch die zentralen Momente $M_{n,m}$ berechnet [212]:

$$M_{n,m} = \sum_{x,y} (\bar{x} - x(u))^n (\bar{y} - y(u))^m \quad (5.3)$$

$$\bar{x} = \frac{1}{N} \sum_{u=0}^{N-1} x(u) \quad \text{und} \quad \bar{y} = \frac{1}{N} \sum_{u=0}^{N-1} y(u) \quad (5.4)$$

(\bar{x}, \bar{y}) bezeichnet den Schwerpunkt der Konturpixel. Die Exzentrizität e_i ist definiert als:

$$e_i = \frac{(M_{2,0} - M_{0,2})^2 + 4 \cdot M_{1,1}}{F} \quad (5.5)$$

$$\alpha_e(i, j) = \frac{|e_i - e_j|}{\max(e_i, e_j)}. \quad (5.6)$$

Die Differenz α_e zweier Konturen bezogen auf die Exzentrizität wird ebenfalls auf das Intervall $[0, 1]$ normiert. Da sich die Exzentrizität aus den Längen der Hauptachsen ableitet, ist sie invariant gegenüber geometrischen Transformationen.

Die beiden globalen Konturdeskriptoren Kompaktheit und Exzentrizität haben den Nachteil, dass durch die starke Aggregation der Konturdaten eine exakte Aussage über die Ähnlichkeit zweier Konturen häufig nicht mehr möglich ist. Für einen ersten Analyseschritt sind die beiden Konturdeskriptoren jedoch gut geeignet, da bei großen Differenzen der Vergleich der Konturen mit komplexeren Klassifikationsverfahren vermieden werden kann.

5.3 Krümmungsbasierter Skalenraum

Eine genauere Aussage über die Ähnlichkeit zweier Konturen ist durch die Analyse ihrer Krümmungen möglich. Besonders stark gekrümmte Bereiche sollen Merkmale zur Beschreibung der Kontur liefern. Bei dem Verfahren des *krümmungsbasierten Skalenraums* wird für jeden Punkt der parametrisierten Kontur die Krümmung berechnet [342, 361, 363, 365]. Nullstellen der Krümmungsfunktion entsprechen Wendepunkten der Kontur, also Übergängen zwischen konkav und konvex gekrümmten Bereichen. Die Kontur wird iterativ durch einen Gaußfilter geglättet, und die Nullstellen der Krümmungsfunktion werden gespeichert. Stark konkav gekrümmte Bereiche einer Kontur bleiben bei der Glättung besonders lang erhalten, so dass die eine konkave Region einschließenden Nullstellen der Krümmungsfunktion als Merkmal zur Beschreibung eines Objektes verwendet werden können [343, 360, 362].

Bei dem Verfahren des krümmungsbasierten Skalenraums handelt es sich um eine Abbildung geometrisch invarianter Faktoren [1, 248, 261]. Die Faktoren sind hier die Nullstellen der Krümmungsfunktion, die während der Glättung (*Evolution*) der Kontur berechnet werden. Die äußere Kontur eines Objektes ist definiert als geschlossene planare Kurve $\Gamma(u)$ mit normalisierter Bogenlänge u , für die gilt:

$$\Gamma(u) = \{(x(u), y(u)) | u \in [0, 1]\}. \quad (5.7)$$

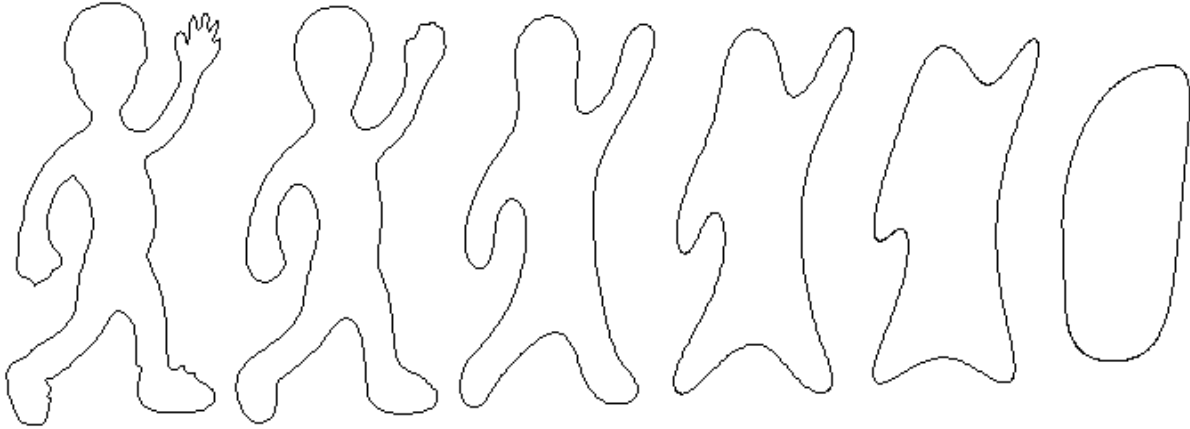


Abbildung 5.2: Glättung einer Kontur mit einem Gaußfilter nach 0, 15, 100, 250, 500 und 2500 Iterationen.

Die Kurve wird mehrfach durch eine eindimensionale Gaußfunktion $g(u, n)$ mit einer Standardabweichung σ geglättet. Die Anzahl der Iterationen bzw. die Anzahl der Glättungen der Kontur wird mit n bezeichnet. In der geglätteten Kurve $\Gamma(u, n)$ beschreibt $(X(u, n), Y(u, n))$ die Position eines Konturpixels $(x(u), y(u))$ nach der Glättung mit der Gaußfunktion. Abbildung 5.2 verdeutlicht die Glättung einer Kontur.

Die Krümmung in einem Punkt der Kontur nach n Iterationen des Glättungsprozesses wird durch die ersten und zweiten Ableitungen $X_u(u, n)$, $Y_u(u, n)$, $X_{uu}(u, n)$ und $Y_{uu}(u, n)$ an der Position u berechnet [364]:

$$\kappa(u, n) = \frac{X_u(u, n) \cdot Y_{uu}(u, n) - X_{uu}(u, n) \cdot Y_u(u, n)}{(X_u(u, n)^2 + Y_u(u, n)^2)^{3/2}}. \quad (5.8)$$

5.4 Abbildungen im krümmungsbasierten Skalenraum

Als Merkmale zur Beschreibung und Klassifikation eines Objektes werden die Wendepunkte der Kontur während der Glättung betrachtet. Eine *Abbildung im krümmungsbasierten Skalenraum* (engl. *curvature scale space image*) bildet die Wendepunkte während des Glättungsprozesses ab, die den Nullstellen der Krümmungsfunktion ($\kappa(u, n) = 0$) entsprechen. Eine Abbildung im krümmungsbasierten Skalenraum ist definiert als:

$$I(u, n) = \{(u, n) | \kappa(u, n) = 0\}. \quad (5.9)$$

Die krümmungsbasierte Skalenraumabbildung kann als Binärbild dargestellt werden, in dem

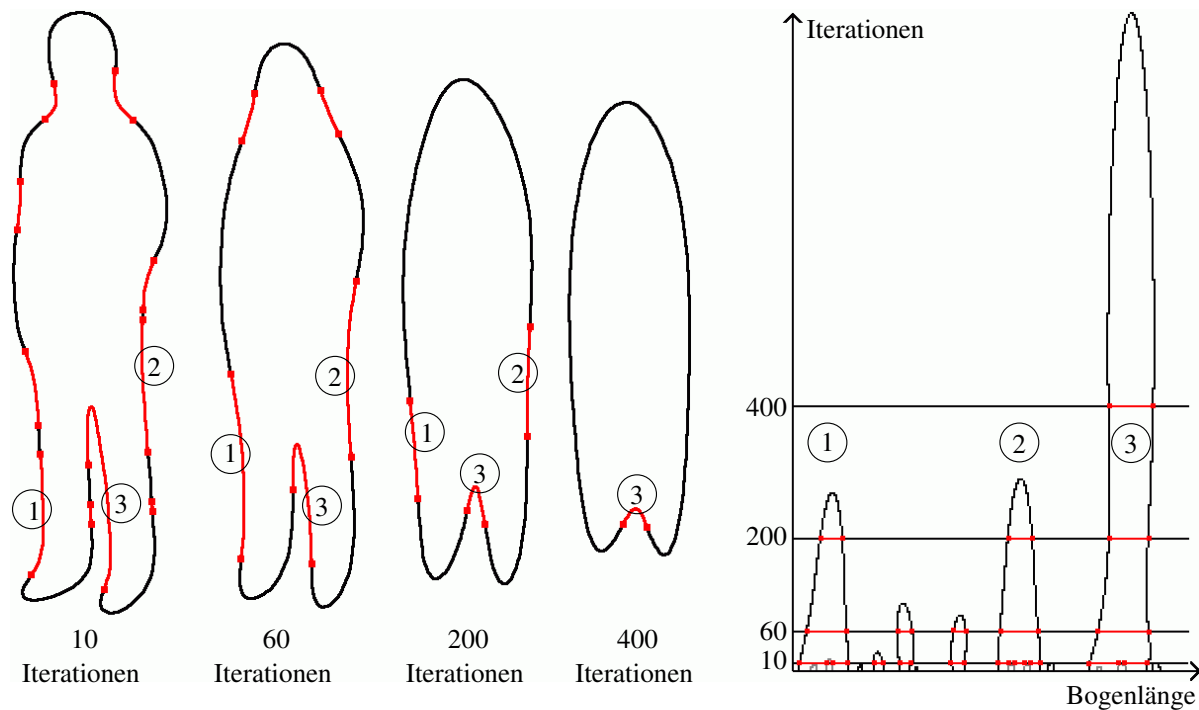


Abbildung 5.3: Glättung einer Kontur nach 10, 60, 200 und 400 Iterationen. Die Wendepunkte der Krümmungsfunktion werden durch Punkte auf den Konturen hervorgehoben. Auf der rechten Seite ist die entsprechende Abbildung im krümmungsbasierten Skalenraum dargestellt. Drei ausgeprägte konvexe Bereiche sind in den Konturen markiert und entsprechen den Bögen der Skalenraumabbildung.

die Wendepunkte der Kontur markiert sind. Abbildung 5.3 zeigt eine Kontur während der Glättung und das entsprechende Skalenraumbild. Auf der horizontalen Achse im Skalenraumbild ist die Position des Pixels auf der Kontur durch die Bogenlänge u festgelegt, die vertikale Achse definiert die Anzahl der Iterationen der Gaußglättung. Jeder Punkt im krümmungsbasierten Skalenraumbild markiert einen Wendepunkt der Krümmung der Kontur an der Position u und der Iteration n .

Während des Glättungsprozesses konvergiert die Kontur gegen einen kreisförmigen Punkt [154, 177], so dass alle konkaven Bereiche verschwinden. Dabei nähern sich jeweils zwei Wendepunkte, die einen konkaven Bereich einschließen, einander an. Deutlich ausgeprägte konkave Bereiche bleiben auch nach vielen Iterationen während des Glättungsprozesses erhalten und werden durch einen hohen Bogen im krümmungsbasierten Skalenraumbild repräsentiert. Das Maximum eines Bogens gibt die Position auf der Kontur und die Anzahl der Iterationen der Gaußglättung an, in der der konkave Bereich gerade noch nicht geglättet ist. Die Höhe eines Bogens im krümmungsbasierten Skalenraumbild steigt mit der Länge des konkaven Bereiches

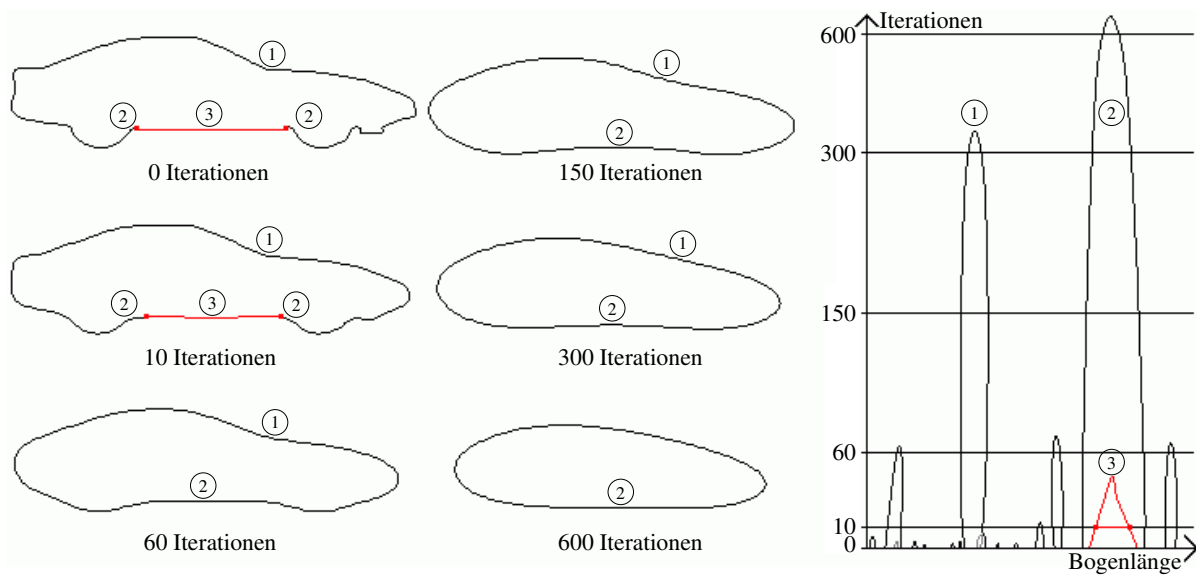


Abbildung 5.4: Glättung einer Kontur nach 0, 10, 60, 150, 300 und 600 Iterationen. Eine lange konvexe Region (3) wird durch zwei stark konkav gekrümmte Bereiche (2) eingeschlossen und erscheint in der Skalenraumabbildung als Bogen.

und der Stärke der Krümmung.

Wird ein konvexer Bereich der Kontur durch zwei stark konkav gekrümmte Bereiche eingeschlossen, so kann dieser als Bogen im Skalenraumbild erscheinen. Abbildung 5.4 verdeutlicht, dass innerhalb weniger Iterationen zunächst die inneren konvexen Bereiche geglättet werden (Abb. 5.4 (3)). Die beiden benachbarten konkaven Bereiche fallen zusammen und ergeben einen großen konkaven Bereich (Abb. 5.4 (2)), der als stark ausgeprägter Bogen oberhalb der kleineren Bögen im Skalenraumbild liegt.

5.5 Vergleich von Konturen

Zur Berechnung der Ähnlichkeit zweier Konturen werden die lokalen Maxima des krümmungsbasierten Skalenraumbildes als Merkmalspunkte ermittelt und miteinander verglichen. Bei einem Vergleich zweier Konturen werden nur konkave Regionen berücksichtigt, da Bögen, die konvexe Regionen beschreiben, immer von einem wesentlich stärker ausgeprägten konkaven Bogen eingeschlossen sind.

Geringe Änderungen einer Kontur, die durch Rauschen bzw. eine ungenaue Segmentierung verursacht werden können, sollten keinen großen Einfluss auf die Merkmale im Skalenraumbild haben. Dies wird durch das Verfahren implizit gewährleistet, da kleine konkave oder

konvexe Bereiche nach wenigen Iterationen geglättet werden, so dass sich als Merkmale zur Beschreibung einer Kontur alle Bögen im Skalenraumbild eignen, die eine Mindesthöhe überschreiten. Schwache Änderungen einer Kontur haben nur sehr geringe Auswirkungen auf die Merkmalspunkte der Skalenraumabbildungen. Ein weiterer Vorteil liegt darin, dass wenige Bögen zur Beschreibung einer Kontur ausreichen. Werden beispielsweise in Abbildung 5.3 alle Wendepunkte und somit alle Bögen ignoriert, die innerhalb der ersten sechzig Iterationen geglättet werden, so bleiben fünf Bögen zur Beschreibung der Kontur erhalten.

Jeder Bogen wird durch die beiden ganzzahligen Werte *Position* und *Höhe* charakterisiert. Die Höhe entspricht der maximalen Höhe des Bogens und gibt die Anzahl der Iterationen an, bei denen die Wendepunkte gerade noch nicht geglättet sind. Die Position des Maximums ermöglicht eine Aussage über die relative Position des konkaven Bereiches zu anderen konkaven Bereichen der Kontur.

Mit einem *Greedy*-Verfahren [19, 92] wird – beginnend mit dem höchsten Bogen – jeder Bogen der ersten Skalenraumabbildung ausgewählt und einem passenden Bogen der zweiten Skalenraumabbildung zugeordnet [429]. $P_1(i) = (u_i, n_i)$ bezeichnet die Position des Maximums des i -ten Bogens der ersten Abbildung, $P_2(j) = (u_j, n_j)$ einen beliebigen Bogen im zweiten Skalenraumbild. Eine *Zuordnung zweier Bögen* P_1 und P_2 ist nur dann möglich, wenn folgende Bedingungen erfüllt sind:

$$D_H(i, j) := |n_i - n_j| < T_H \quad (5.10)$$

$$D_P(i, j) := \min(|u_i - u_j|, N - |u_i - u_j|) < T_P. \quad (5.11)$$

Die Differenz der Höhe beider Bögen D_H darf einen Schwellwert T_H nicht übersteigen, da sonst die Unterschiede der konkaven Bereiche zu groß und die beiden Bereiche des Objektes nicht mehr vergleichbar sind. Zusätzlich dürfen die Positionen der Bögen nicht allzu stark voneinander abweichen. Die abgetastete Kontur wird durch N Konturpixel beschrieben, wobei in der geglätteten Kontur die Pixel an den Positionen 0 und $N - 1$ benachbart sind. Der maximale Abstand zwischen beliebigen Punkten – gemessen in der Anzahl der Konturpixel – kann maximal $\frac{N}{2}$ betragen. Das Minimum aus $|u_i - u_j|$ und $N - |u_i - u_j|$ gibt die tatsächliche Entfernung für zwei beliebige Positionen u_i und u_j an und muss unter dem Schwellwert T_P liegen, damit beide Bögen als ähnlich gelten.

Als Differenz $D(i, j)$ zweier *ähnlicher Bögen* i und j wird die euklidische Distanz aus Positions- und Höhendifferenzen berechnet, welche ein kompaktes Maß für die visuelle Ähnlichkeit

zweier konkaver Bereiche einer Kontur liefert [364]:

$$D(i, j) = \begin{cases} \sqrt{D_H^2(i, j) + D_P^2(i, j)} & \text{falls } D_H(i, j) < T_H \text{ und} \\ & D_P(i, j) < T_P \\ F \cdot \max(n_i, n_j) & \text{sonst.} \end{cases} \quad (5.12)$$

Statt zwei einzelne Distanzen zu berechnen, werden in $D(i, j)$ sowohl Abweichungen der Position einer konkaven Region als auch Unterschiede bezüglich der Stärke der Krümmung kombiniert. Wird die maximal zulässige Höhen- oder Positionsdivergenz überschritten, so können beide Bögen nicht miteinander verglichen werden, und als Differenz wird die mit einem Faktor F gewichtete Höhe des größeren Bogens festgelegt. Die Summe der Differenzen aller Bögen beschreibt die Ähnlichkeit zweier Abbildungen im krümmungsbasierten Skalenraum.

5.5.1 Rotationsinvarianter Konturvergleich

Die Wahl eines anderen Startpunktes bei der Abtastung der Kontur verschiebt das Skalenraumbild in horizontaler Richtung. Gleiches gilt für eine Rotation eines Objektes, die mit einem geänderten Startpunkt vergleichbar ist. Um Rotationsinvarianz bei einem Vergleich zweier Skalenraumbilder zu gewährleisten, werden die Bögen von einem der beiden Skalenraumbilder horizontal entlang der x-Achse verschoben. Die aus dem Bild hinausgeschobenen Bögen erscheinen an der gegenüberliegenden Seite der Skalenraumabbildung.

Eine Umkehrung der Abtastrichtung hat die gleiche Auswirkung auf das Skalenraumbild wie eine Spiegelung der Kontur. Beide erzeugen ein an der y-Achse gespiegeltes Skalenraumbild. Eine Rotation oder Spiegelung der Kontur wird daher durch eine horizontale Verschiebung bzw. Spiegelung der Bögen ausgeglichen.

Um Rotationen zu kompensieren, werden vor dem Vergleich zweier Skalenraumabbildungen die Positionen der Bögen passend ausgerichtet [294, 429]. Die Positionen der k größten Bögen beider Skalenraumabbildungen werden in zwei Listen gespeichert. Für alle Kombinationen der Elemente beider Listen wird ein Vektor ermittelt, der angibt, wie weit das erste Skalenraumbild verschoben werden muss, damit die beiden ausgewählten Bögen an derselben Position liegen. Für alle Verschiebungsvektoren werden die Differenzen der Skalenraumabbildungen berechnet. Die minimale Differenz entspricht der besten Rotation und definiert die Ähnlichkeit beider Konturen.

Um eine gespiegelte Kontur zu erkennen, wird das Skalenraumbild der ersten Kontur an der y-Achse gespiegelt und die Differenz erneut berechnet. Die gespiegelte Position u'_i eines Bogens

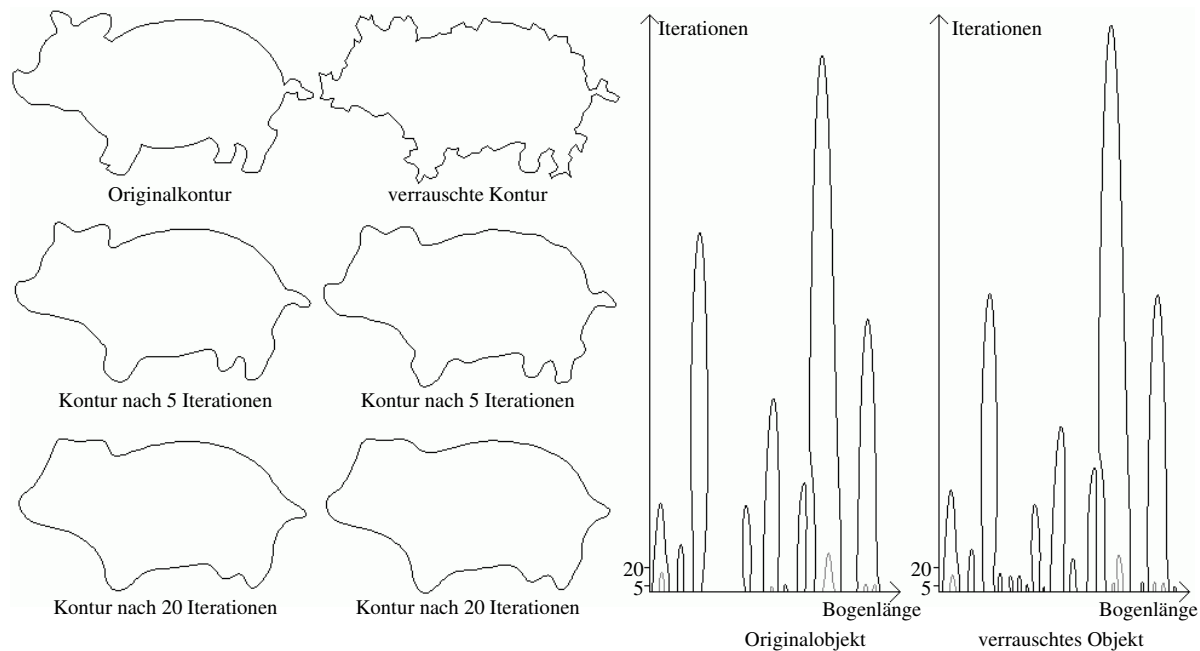


Abbildung 5.5: Auswirkung von Rauschen auf Abbildungen im krümmungsbasierten Skalenraum: Originalkontur und verrauschte Kontur nach 0, 5 und 20 Iterationen (links). Skalenraumabbildungen beider Objekte (rechts).

entspricht im Skalenraumbild der Spiegelung an der y-Achse und wird durch $u'_i = N - u_i$ berechnet.

5.5.2 Merkmale der Abbildungen im krümmungsbasierten Skalenraum

Das vorgestellte Verfahren zur Klassifikation von Konturen weist eine Vielzahl positiver Eigenschaften auf. Komplexe Konturen können mit wenigen Wertepaaren beschrieben werden, so dass nur wenige Daten gespeichert werden müssen. Der Aufwand für die Berechnung der Differenz zweier Abbildungen im krümmungsbasierten Skalenraum ist relativ gering, da nur die euklidischen Distanzen weniger Wertepaare summiert werden müssen. Durch die Ausrichtung und Spiegelung der Bögen wird das Verfahren rotationsinvariant, so dass keine gedrehten oder gespiegelten Objekte als Referenzobjekte in eine Datenbank eingefügt werden müssen. Die Objektgröße bzw. die Skalierung eines Bildes hat nur geringe Auswirkungen auf die Skalenraumabbildung, da alle Objekte mit einer festen Anzahl von Konturpixeln abgetastet werden.

Ein weiterer Vorteil ist die Unempfindlichkeit gegenüber Rauschen und kleineren Bildfehlern. Abbildung 5.5 zeigt die Originalkontur und eine verrauschte Kontur mit den entsprechenden

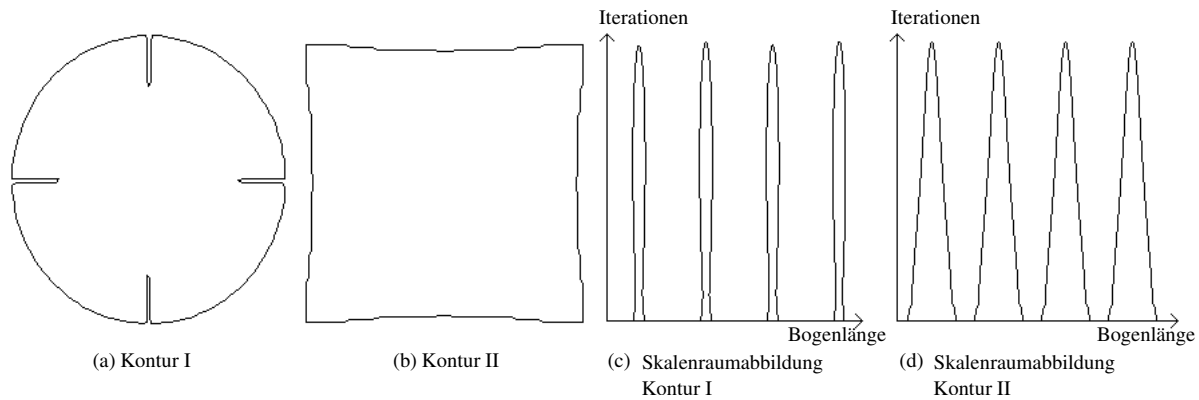


Abbildung 5.6: Zwei unterschiedliche Konturen können sehr ähnliche Skalenraumabbildungen erzeugen. Anhand der Position und Höhe der Bögen gelten beide Konturen als identisch.

Abbildungen im krümmungsbasierten Skalenraum. Schon innerhalb der ersten Iterationen des Glättungsprozesses wird ein großer Teil des Rauschens aus der Kontur entfernt, die stark ausgeprägten konkaven Bereiche bleiben dagegen in beiden Abbildungen erhalten.

Den genannten Vorteilen stehen zum Teil sehr ungenaue Klassifikationsergebnisse gegenüber. Zwei von uns neu entwickelte Verfahren zur Verbesserung der Ergebnisse des ursprünglichen Verfahrens werden in den folgenden beiden Abschnitten vorgestellt.

5.6 Vermeidung von Mehrdeutigkeiten

Die Bögen einer Skalenraumabbildung beschreiben die Ausprägungen und relativen Positionen konkaver Bereiche einer Kontur [2]. Die Länge und die Stärke einer Krümmung wird durch die Höhe des Bogens charakterisiert. Das Beispiel in Abbildung 5.6 verdeutlicht, dass zwei unterschiedliche Konturen sehr ähnliche Abbildungen im krümmungsbasierten Skalenraum erzeugen können, in denen die Positionen und Höhen der Bögen nahezu identisch sind. Es wird im Folgenden ein neues von uns entwickeltes Verfahren vorgeschlagen, um zu verhindern, dass signifikant unterschiedliche konkave Regionen zu nahezu identischen Merkmalswerten in Skalenraumabbildungen führen [429]. Die Länge eines konkaven Bereiches entspricht im krümmungsbasierten Skalenraumbild der Breite des Bogens der Originalkontur vor der ersten Glättung. Zur Vermeidung von Mehrdeutigkeiten wird neben der *Position* und *Höhe* eines Bogens auch dessen *Breite* als Merkmal berücksichtigt.

Die Differenz zweier Bögen i und j wird unter Berücksichtigung der Länge der konkaven

Bereiche wie folgt berechnet:

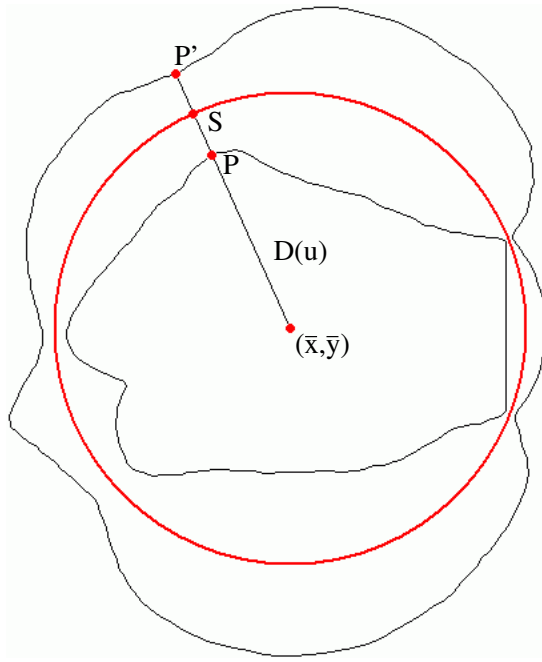
$$D(P_i, P_j) = \begin{cases} \sqrt{D_H^2 + D_P^2 + D_W^2} & \text{falls } D_H < T_H \text{ und} \\ & D_P < T_P \text{ und} \\ & D_W < T_W \\ F \cdot \max(n_i, n_j) & \text{sonst.} \end{cases} \quad (5.13)$$

D_H , D_P und D_W bezeichnen die absoluten Differenzen in Bezug auf Höhe, Position und Breite zweier Bögen. Analog zu der Position oder Höhe verhindern deutliche Unterschiede in der Breite die Zuordnung und den Vergleich der beiden Bögen. Da die Höhe sowohl die Länge als auch die Stärke der Krümmung wiedergibt, bestimmt ausschließlich die gewichtete Höhe des größeren Bogens die Differenz für zwei deutlich unterschiedliche Bögen.

5.7 Klassifikation konvexer Objektregionen

In diesem Abschnitt wird ein neues Verfahren vorgestellt, um Merkmale zur Beschreibung konvexer Objektregionen zu ermittelt. Konvexe Bereiche einer Kontur werden nur unzureichend berücksichtigt und haben sehr geringe Auswirkungen auf eine Abbildung im krümmungsbasierten Skalenraum, da ein konvexer Bereich während der Glättung nur indirekt die beiden angrenzenden konkaven Bereiche beeinflusst und diese die Position und Höhe der Bögen bestimmen. So glättet eine stark konvex gekrümmte Region im Vergleich zu einer schwach konvexen den benachbarten konkaven Bereich schneller. Konvexe Objekte – also Objekte ohne konkave Regionen – können anhand ihrer Abbildungen im krümmungsbasierten Skalenraum nicht unterschieden werden. Eine geometrische Figur heißt dann *konvex*, wenn für zwei beliebige Punkte dieser Figur alle Punkte der Verbindungsstrecke zur Fläche der Figur gehören [45].

Im Allgemeinen werden konvexe Bereiche einer Kontur während der Glättung nicht durch Wendepunkte eingeschlossen, so dass aus den Bögen im krümmungsbasierten Skalenraumbild keine Rückschlüsse auf konvexe Bereiche gezogen werden können. Um dennoch Merkmale für diese Bereiche zu erhalten, wird eine neue Kontur erstellt, die als *transformierte Kontur* bezeichnet wird. Durch die Transformation werden stark konvex gekrümmte Bereiche in konkave Bereiche umgewandelt und umgekehrt. Eine Möglichkeit zur Erzeugung einer transformierten Kontur ist die Spiegelung der Konturpixel an einem Kreis, der um die Kontur gelegt wird. Abbildung 5.7 verdeutlicht die Transformation einer Kontur durch Spiegelung der Konturpixel an einer Kreislinie.



(\bar{x}, \bar{y}) : Der Mittelpunkt des Kreises entspricht dem Schwerpunkt der Kontur

P: Pixel $(x(u), y(u))$ der ursprünglichen Kontur

$D(u)$: Abstand zwischen (\bar{x}, \bar{y}) und $(x(u), y(u))$

S: Punkt der Kreislinie an dem $(x(u), y(u))$ gespiegelt wird

P': Gespiegeltes Konturpixel $(x'(u), y'(u))$ der transformierten Kontur

Abbildung 5.7: Transformation einer Kontur

Der Schwerpunkt (\bar{x}, \bar{y}) aller Konturpixel $(x(u), y(u))$ wird entsprechend Gleichung 5.4 als Mittelpunkt des Kreises festgelegt. Der Radius R des Kreises wird so gewählt, dass alle Konturpixel auf der Kreisfläche liegen:

$$R = \max_u \left\{ \sqrt{(\bar{x} - x(u))^2 + (\bar{y} - y(u))^2} \right\}. \quad (5.14)$$

Jedes Konturpixel $(x(u), y(u))$ wird entlang der Geraden durch (\bar{x}, \bar{y}) und $(x(u), y(u))$ an der Kreislinie im Punkt S gespiegelt. Der Abstand $D(u)$ des Punktes $(x(u), y(u))$ zum Mittelpunkt des Kreises beträgt:

$$D_u = \sqrt{(\bar{x} - x(u))^2 + (\bar{y} - y(u))^2}. \quad (5.15)$$

Die Entfernung des gespiegelten Punktes $(x'(u), y'(u))$ durch die Spiegelung an der Kreislinie beträgt zum Mittelpunkt $D(u) + 2 \cdot (R - D(u)) = 2R - D(u)$. Abbildung 5.8 verdeutlicht, dass mit Hilfe des Strahlensatzes folgende Beziehung abgeleitet werden kann:

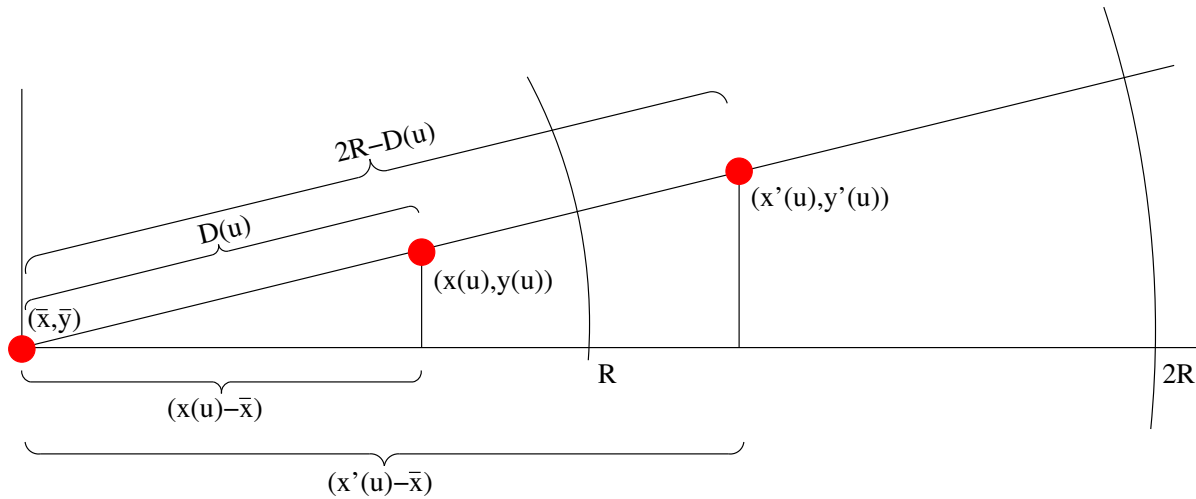


Abbildung 5.8: Berechnung der Position eines transformierten Konturpixels mit Hilfe des Strahlensatzes

$$\frac{2R - D(u)}{D(u)} = \frac{x'(u) - \bar{x}}{x(u) - \bar{x}}. \quad (5.16)$$

Die analoge Beziehung gilt für die y-Koordinate. Durch Umformung der Gleichung 5.16 wird die Position $(x'(u), y'(u))$ ermittelt:

$$x'(u) = \frac{2R - D(u)}{D(u)} \cdot (x(u) - \bar{x}) + \bar{x} \quad (5.17)$$

$$y'(u) = \frac{2R - D(u)}{D(u)} \cdot (y(u) - \bar{y}) + \bar{y} \quad (5.18)$$

Jedes abgetastete Konturpixel wird entlang der Geraden durch (\bar{x}, \bar{y}) und $(x(u), y(u))$ an der Kreislinie gespiegelt. Entspricht die Krümmung in der lokalen Umgebung von $(x(u), y(u))$ der Krümmung der Kreislinie, so bleibt die Stärke der Krümmung des entsprechenden Bereiches in der transformierten Kontur nahezu unverändert. Konvexe Bereiche der Kontur, die stärker als die Kreislinie gekrümmt sind, ergeben konkave Bereiche in der transformierten Kontur. Umgekehrt wird ein konkav gekrümmter Bereich in einen stark konvex gekrümmten Bereich transformiert. Der Zusammenhang zwischen stark konvex gekrümmten Bereichen der ursprünglichen Kontur und konkaven Bereichen der transformierten Kontur wird in Abbildung 5.7 deutlich.



Abbildung 5.9: Von einem Punkt M werden entlang einer Geraden die Schnittpunkte mit der Kontur gezählt. Bei einer ungeraden Anzahl an Schnittpunkten liegt der Punkt M innerhalb der Objektes.

Statt die Kontur an einer Kreislinie zu spiegeln, könnten auch andere geometrische Formen ausgewählt werden. Eckige Figuren haben jedoch den Nachteil, dass das transformierte Objekt in der Nähe einer Ecke stark durch die Ecke beeinflusst wird und die Krümmung in diesen Bereichen nicht kontinuierlich verläuft. Auch Figuren ohne Ecken wie beispielsweise Ellipsen eignen sich nur eingeschränkt zur Erzeugung von transformierten Konturen. Um Invarianz gegenüber Rotationen zu erhalten, müssten die Hauptachsen der Kontur und der Ellipse passend ausgerichtet werden, was mit einem zusätzlichen Rechenaufwand verbunden ist. Zudem wäre das Verhältnis der Länge der Hauptachse zur Nebenachse für jedes Objekt neu zu bestimmen. Bei der Erzeugung der transformierten Kontur durch Spiegelung an einem Kreis werden die aufgeführten Probleme vermieden.

Liegt ein Konturpixel genau auf dem Kreismittelpunkt, so ist die Richtung, in der dieser Punkt gespiegelt werden soll, nicht definiert. Zwei Lösungen bieten sich an, um den transformierten Punkt zu bestimmen. Die transformierte Position eines Pixels kann durch Interpolation der benachbarten transformierten Pixel berechnet werden. Alternativ ist eine geringe Verschiebung des Kreismittelpunktes möglich. Der Mittelpunkt des Kreises sollte dabei so verschoben werden, dass er nach der Verschiebung innerhalb des Objektes liegt. Zur Überprüfung, ob ein Punkt innerhalb oder außerhalb der Kontur liegt, wird eine Gerade durch den Punkt gelegt. Ausgehend vom Punkt in eine beliebige Richtung entlang der Geraden werden die Schnittpunkte mit der Kontur gezählt. Bei einer ungeraden Anzahl an Schnittpunkten liegt der Punkt innerhalb des Objektes, bei einer geraden Anzahl außerhalb, wobei Berührungspunkte nicht als Schnittpunkte gelten. Abbildung 5.9 verdeutlicht, wie durch beliebige Geraden geprüft werden kann, ob ein Punkt innerhalb oder außerhalb einer Kontur liegt.

Durch Rauschen ändern sich die Positionen einzelner Konturpixel, so dass die Größe des umgebenden Kreises variieren kann. Der gewählte Radius hat jedoch nur eine geringe Auswirkung auf die transformierte Kontur. Im Wesentlichen treten Skalierungsunterschiede auf, die im Skalenraumbild nicht abgebildet werden, so dass Rauschen sowohl im ursprünglichen als auch im transformierten Skalenraumbild nur einen geringen Einfluss hat.

5.8 Aggregation der Klassifikationsergebnisse für Videosequenzen

Es wird die Annahme getroffen, dass ein Objekt innerhalb einer Kameraeinstellung in mehreren Bildern hintereinander sichtbar ist. Da in einzelnen Bildern sowohl Fehler bei der Segmentierung als auch bei der Klassifikation auftreten können, werden die Ergebnisse aggregiert, um einzelne fehlerhafte Ergebnisse zu eliminieren. Es wird davon ausgegangen, dass sich nur ein Objekt im Bild bewegt bzw. dass bei mehreren Objekten jeweils dasselbe Objekt durch Analyse der Größe und Position der segmentierten Bereiche ausgewählt wird. Neben der Aggregation über die Anzahl der erkannten Objektklassen wird ein neues Verfahren vorgestellt, das die Distanz zwischen Objekt und Objektklasse berücksichtigt. Ein Maß für die Zuverlässigkeit wird eingeführt, durch das die Verlässlichkeit eines Klassifikationsergebnisses spezifiziert wird.

5.8.1 Aggregation über die Anzahl der erkannten Objektklassen

Für jedes Bild i ($i = 1 \dots N$) der Kameraeinstellung wird das ähnlichste Objekt j ($j = 1 \dots M$) der Datenbank und der entsprechende Name der Objektklasse ermittelt. Der Aufbau der Datenbank mit den verfügbaren Objekten und Objektklassen wird im Rahmen der experimentellen Ergebnisse in Abschnitt 5.9.1 vorgestellt.

Der Name des in der Kameraeinstellung dargestellten Objektes wird definiert als Name der am häufigsten erkannten Objektklasse. Der relative Unterschied zwischen der Objektklasse mit der größten und zweitgrößten Anzahl an erkannten Objekten liefert ein *Maß für die Zuverlässigkeit einer korrekten Klassifikation*. Für jede Objektklasse k wird der relative Anteil der erkannten Bilder mit R_k bezeichnet. Die Objektklassen werden anhand ihrer relativen Anteile absteigend sortiert, so dass gilt: $R_1 \geq R_2 \dots \geq R_K$. Die Zuverlässigkeit β_R für eine korrekte Klassifikation des Objektes der Kameraeinstellung wird definiert als:

$$\beta_R = \frac{2 \cdot R_1}{R_1 + R_2} - 1 \in [0, 1]. \quad (5.19)$$

Werden ähnlich viele Objekte den Objektklassen R_1 und R_2 zugeordnet, so liegt β_R nahe bei null. Der Wert steigt mit wachsenden Differenzen zwischen R_1 und R_2 bis zu dem maximalen Wert von eins, bei dem alle Objekte einer Objektklasse zugeordnet sind.

5.8.2 Aggregation über die Distanz zur Objektklasse

Die *Distanz* $d_{k,i}$ zwischen *Objekt* i und *Objektklasse* k ist definiert als das Minimum der Distanzen $D_{j,i}$ zwischen dem Objekt und allen Objekten j der Objektklasse k . Treten in einzelnen Bildern einer Videosequenz Segmentierungsfehler auf oder gibt es Objekte, die durch keine Objektklasse in der Datenbank repräsentiert werden, so ist es notwendig, die Distanzen zu allen Objektklassen zu berücksichtigen. Bei fehlerhaft klassifizierten Konturen sind häufig sehr hohe Differenzwerte zu allen Objektklassen zu beobachten. In diesem Fall wird angenommen, dass es sich um ein unbekanntes Objekt handelt.

Überschreitet die Differenz zwischen dem unbekannten Objekt und einem Objekt der Datenbank einen Schwellwert T_d , so bleibt das Klassifikationsergebnis für dieses Objekt unberücksichtigt. $d_{k,i}$ speichert für eine Kontur i die minimale Distanz zur Objektklasse k . Das Ähnlichkeitsmaß C_k beschreibt in aggregierter Form, wie ähnlich alle Objekte einer Kameraeinstellung einer Objektklasse k sind. Die minimalen Distanzen zur Objektklasse k gehen umgekehrt proportional in C_k ein:

$$C_k = \frac{1}{N} \sum_{i=1}^N \begin{cases} \frac{1}{1+d_{k,i}} & \text{falls } d_{k,i} < T_d \\ 0 & \text{sonst.} \end{cases} \quad (5.20)$$

Existiert für jedes Objekt ein identisches Objekt in der Datenbank, so liegen die Distanzen $d_{k,i}$ bei null, und C_k erhält einen maximalen Wert von eins. Wird in der Sequenz kein ähnliches Objekt gefunden ($d_{k,i} \geq T_d$), so erhält C_k den Wert von null. Das Maximum von C_k spezifiziert die abgebildete Objektklasse k der Videosequenz, wobei die Zuverlässigkeit der Klassifikation entsprechend Gleichung 5.19 aus der Differenz der beiden größten Werte abgeleitet wird.



Abbildung 5.10: Beispielobjekte der Datenbank aus der Objektklasse PKW

5.9 Experimentelle Ergebnisse

Im Rahmen der experimentellen Ergebnisse werden die Datenbank mit den Referenzobjekten und die analysierten Videosequenzen vorgestellt. Die Erkennung der Objekte erfolgt durch einen Vergleich der aus den Skalenraumabbildungen ermittelten Merkmale. Abschließend werden die Klassifikationsergebnisse der neu entwickelten Verfahren vorgestellt.

5.9.1 Objekte der Datenbank

Die Objekte der Datenbank sind in sechs Objektklassen eingeteilt: *Säugetier*, *Vogel*, *PKW*, *Person*, *Flugzeug* und *Schiff*. Für die Datenbank wurden typische und leicht erkennbare Ansichten eines Objektes (*kanonische Sichten*) ausgewählt, die es einem Menschen ermöglichen, das Objekt besonders schnell und zuverlässig zu erkennen.

Die Hälfte der Objekte der Datenbank stammt aus einer Clipart-Bibliothek und enthält schematische Zeichnungen. Der andere Teil wurde automatisch aus Videosequenzen segmentiert, so dass diese Objekte typische Segmentierungsfehler – wie beispielsweise den Schatten eines Objektes – enthalten. Abbildung 5.10 zeigt exemplarisch einzelne Objekte der Datenbank aus der Objektklasse *PKW*. Tabelle 5.1 gibt die Anzahl und Verteilung der Objekte innerhalb der Datenbank an.

Jede Objektklasse wird durch 22 bis 137 Objekte repräsentiert. Insbesondere in der Objektklasse *Person* sind überdurchschnittlich viele Objekte enthalten, da eine Kontur sehr stark von der Position der Arme und Beine abhängt und sie sich im Vergleich zu Konturen starrer Objekte innerhalb kurzer Zeit deutlich ändern kann. Um eine zuverlässige Erkennung der Objektklasse *Säugetier* zu ermöglichen, wäre eine deutlich größere Anzahl von Objekten erforderlich. Obwohl eine zuverlässige Erkennung wegen der zu geringen Anzahl nicht möglich ist, bleiben die Objekte in der Datenbank gespeichert, da sie eine allgemeinere Aussage über die Zuverlässigkeit der Algorithmen zur Objekterkennung ermöglichen.

Alle Objekte der Datenbank haben einen monochromen Hintergrund, so dass die Ermittlung der äußeren Kontur keinen manuellen Eingriff eines Benutzers erfordert. Nach der Parame-

Name der Objektklasse	Anzahl der Elemente in der Datenbank	Durchschnitt Kompaktheit (Varianz)	Durchschnitt Exzentrizität (Varianz)
Säugetier	38	6,4 (2,2)	1,7 (0,5)
Vogel	25	5,6 (1,9)	1,9 (0,8)
Flugzeug	22	6,5 (3,7)	2,8 (1,6)
Schiff	27	3,0 (1,0)	2,1 (0,8)
PKW	63	2,0 (0,5)	2,0 (0,6)
Person	137	5,2 (2,7)	2,7 (0,9)
Summe / Durchschnitt	312	4,6 (2,7)	2,3 (0,9)

Tabelle 5.1: Verteilung der Objekte der Datenbank auf die Objektklassen

trisierung der Kontur wird jede durch genau 200 gleichmäßig auf der Kontur verteilte Punkte in Form von Wertepaaren beschrieben. 200 Konturpixel bieten einen guten Kompromiss zwischen den erfassten Details einer Kontur und dem Rechenaufwand, da jede Verdopplung der Konturpixel die Anzahl der benötigten Iterationen für die Glättung der Kontur zur Erzeugung der Skalenraumabbildung ungefähr um den Faktor vier erhöht. Für jede Kontur ist der Name der Objektklasse definiert. Zusätzlich werden die globalen Konturdeskriptoren *Kompaktheit* und *Exzentrizität* berechnet und gespeichert.

Tabelle 5.1 gibt für jede Objektklasse den Durchschnitt und die Varianz für die globalen Konturdeskriptoren an. Ein niedriger Wert für die Kompaktheit bedeutet eine hohe Ähnlichkeit mit einem Kreis; er tritt insbesondere in der Objektklasse *PKW* auf. Die Exzentrizität beschreibt die Verteilung der Konturpixel entlang der Hauptachsen und erreicht besonders hohe Werte bei Personen und Flugzeugen. Besonders hohe Varianzen treten bei den globalen Deskriptoren innerhalb der Klasse *Flugzeug* auf, da diese aus unterschiedlichen Perspektiven aufgenommen werden und nur in einem Teil der Konturen Flügel sichtbar sind.

Für alle Objekte der Datenbank werden die Skalenraumabbildungen berechnet und deren *relevante Bögen* ermittelt. Ein Bogen gilt als relevant, falls die Höhe einen Wert von dreißig überschreitet, d. h. dass der entsprechende konkave Bereich der Kontur nicht innerhalb der ersten dreißig Iterationen geglättet wird. Zur Charakterisierung eines Bogens dient dessen Höhe, Position und Breite. Die Berechnung der Merkmale für alle Objekte der Datenbank benötigt weniger als fünf Minuten Rechenzeit auf einem durchschnittlich leistungsfähigen PC. Zu diesen Rechenschritten zählt die Abtastung und Parametrisierung der Kontur, die Berechnung der globalen Konturdeskriptoren, die Transformation der Kontur, die Erzeugung der Skalenraumabbildungen und die Ermittlung und Speicherung der relevanten Bögen der Skalenraumabbildungen.

5.9.2 Testsequenzen zur Objekterkennung

Anhand dreißig kurzer Videosequenzen mit einer Länge zwischen vier und dreißig Sekunden wird die Qualität der Algorithmen zur Klassifikation von Objekten analysiert. Um zu überprüfen, ob die Klassifikation durch ein spezielles Segmentierungsverfahren negativ beeinflusst wird, erfolgt die Segmentierung der Objekte mit unterschiedlichen Verfahren. Die ersten beiden Sequenzen sind mit einer statischen Kamera aufgenommen, so dass die Differenz zwischen Hintergrundbild und Kamerabild ohne Kompensation der Kamerabewegung berechnet werden kann. Das von Kim und Hwang entwickelte Segmentierungsverfahren, in dem Regionen mit starken Pixeldifferenzen analysiert und aggregiert werden, dient zur Segmentierung dieser beiden Sequenzen [255, 256, 257].

Auch die dritte Sequenz ist mit einer statischen Kamera aufgenommen. Die Segmentierung erfolgt mit dem von Paragios und Deriche vorgestellten Verfahren, bei dem der optische Fluss innerhalb eines Videos mit Hilfe eines statistischen Modells analysiert und aus diesem das segmentierte Objekt ermittelt wird [406, 407, 408, 409]. Die Sequenzen 4 – 15 sind manuell segmentiert, so dass Fehler durch die Segmentierung ausgeschlossen werden können. Die Bildwiederholrate dieser Sequenzen liegt zwischen zwei und acht Bildern pro Sekunde. Die Sequenzen 16 – 30 wurden automatisch mit dem in Kapitel 4.4 vorgestellten Algorithmus segmentiert. Der Schatten eines Objektes, der häufig mit dem Objekt zusammen segmentiert wird, kann bei diesen Sequenzen deutliche Segmentierungsfehler verursachen.

In 17 Testsequenzen sind Personen und in 11 sind PKWs abgebildet, wobei die beiden Sequenzen *PKW-6* und *PKW-7* eine Ausnahme bilden, da sie einen Lieferwagen zeigen, für den nur sehr wenige Referenzobjekte in der Datenbank enthalten sind. In zwei weiteren Sequenzen ist eine Taube segmentiert. Tabelle 5.2 gibt einen Überblick über die verwendeten Segmentierungsverfahren und die Längen der Testsequenzen.

5.9.3 Klassifikation mit Hilfe der Merkmale des krümmungsbasierten Skalenraums

Zunächst wird die Erkennung der Objekte der Testsequenzen mit dem ursprünglichen Skalenraumverfahren analysiert. Zur Charakterisierung eines Objektes werden für jeden Bogen des Skalenraumbildes zunächst nur die beiden Merkmale *Position* und *Höhe* verwendet, so dass globale Konturdeskriptoren oder zusätzliche Informationen wie die Breite eines Bogens zunächst unberücksichtigt bleiben. Tabelle 5.2 gibt einen Überblick über die Klassifikationsergebnisse für die einzelnen Testsequenzen, wobei fehlerhafte Ergebnisse bzw. Ergebnisse mit

Nr.	Sequenz	Segmen- tierungs- verfahren	Anzahl Bilder	Anzahl gültiger Bilder		erkannte Objektklasse		Maß für die Zuverlässigkeit
1	Person-1	autom. [257]	26	26	100 %	Person	100 %	1,00
2	Person-2	autom. [257]	39	39	100 %	Person	97 %	0,95
3	Person-3	autom. [407]	39	39	100 %	Person	62 %	0,45
4	Person-4	manuell	29	29	100 %	Person	76 %	0,76
5	Person-5	manuell	13	13	100 %	Person	69 %	0,80
6	Person-6	manuell	165	165	100 %	Person	59 %	0,53
7	Vogel-1	manuell	15	15	100 %	—	33 %	0,00
8	Vogel-2	manuell	67	66	99 %	PKW	62 %	0,24
9	PKW-1	manuell	32	32	100 %	PKW	100 %	1,00
10	PKW-2	manuell	8	8	100 %	PKW	100 %	1,00
11	PKW-3	manuell	51	51	100 %	PKW	100 %	1,00
12	PKW-4	manuell	19	19	100 %	PKW	100 %	1,00
13	PKW-5	manuell	22	21	95 %	PKW	100 %	1,00
14	PKW-6	manuell	57	42	74 %	PKW	55 %	0,35
15	PKW-7	manuell	14	13	93 %	PKW	77 %	0,54
16	Person-7	autom.	39	39	100 %	Person	64 %	0,47
17	Person-8	autom.	42	42	100 %	Person	88 %	0,85
18	Person-9	autom.	239	239	100 %	Person	76 %	0,73
19	Person-10	autom.	28	28	100 %	Person	64 %	0,64
20	Person-11	autom.	82	82	100 %	Person	44 %	0,11
21	Person-12	autom.	151	150	99 %	Person	54 %	0,45
22	Person-13	autom.	31	31	100 %	Person	39 %	0,41
23	Person-14	autom.	35	35	100 %	Person	60 %	0,56
24	Person-15	autom.	300	300	100 %	Person	70 %	0,73
25	Person-16	autom.	261	261	100 %	Person	70 %	0,65
26	Person-17	autom.	28	28	100 %	Person	43 %	0,33
27	PKW-8	autom.	12	12	100 %	PKW	92 %	0,83
28	PKW-9	autom.	14	14	100 %	PKW	86 %	0,71
29	PKW-10	autom.	10	10	100 %	PKW	100 %	1,00
30	PKW-11	autom.	30	29	97 %	PKW	93 %	0,86
Summe / Durchschnitt			1898	1878	99 %	69 %		0,64

Tabelle 5.2: Klassifikationsergebnisse zur Objekterkennung ohne zusätzliche Optimierungsschritte. Höhere Fehlerraten und unzuverlässige Klassifikationsergebnisse sind fett hervorgehoben.

geringer Aussagekraft hervorgehoben sind.

Um zwei Skalenraumbilder zu vergleichen, muss jedem *signifikanten Bogen* der ersten Abbildung, d. h. jedem Bogen, dessen Höhe mindestens 50 Prozent der Höhe der Skalenraumabbildung erreicht, ein entsprechender Bogen in der zweiten Skalenraumabbildung zugeordnet werden können. Die beiden Schwellwerte $T_P = 30\%$ und $T_H = 30\%$ definieren die maximal zulässigen Positions- und Höhendifferenzen zwischen zwei Bögen [429]. Nur wenn für alle signifikanten Bögen ein entsprechender Bogen in der zweiten Skalenraumabbildung gefunden wird, besteht eine gewisse Ähnlichkeit zwischen beiden Objekten, und der Differenzwert basierend auf der euklidischen Distanz der Maxima der Bögen wird berechnet.

Die Spalte *Anzahl gültiger Bilder* in Tabelle 5.2 gibt an, für wie viele Objekte mindestens ein ähnliches Objekt in der Datenbank gefunden werden konnte. Eine höhere Anzahl ungültiger Bilder tritt nur in der Sequenz *PKW-6* durch den Lieferwagen auf. In der Spalte *erkannte Objektklasse* ist der Name und der prozentuale Anteil der am häufigsten erkannten Objektklasse bezogen auf die Anzahl der gültigen Bilder angegeben. Nur in den beiden Vogelsequenzen wurde eine fehlerhafte bzw. keine Objektklasse spezifiziert. Dies ist insbesondere auf die viel zu geringe Anzahl an Vögeln in der Datenbank zurückzuführen.

Für jede Sequenz wird das Maß für die Zuverlässigkeit entsprechend der Gleichung 5.19 berechnet. In der Sequenz *Vogel-1* ist dieser Wert null, so dass zwei Objektklassen gleich viele Objekte zugeordnet werden und die korrekte Objektklasse somit nicht erkannt wird. Es wird angenommen, dass bei Werten von mindestens 0,6 die Klassifikation mit hoher Wahrscheinlichkeit korrekt ist. Entsprechend dieses Wertes werden neun PKW-Sequenzen und neun Sequenzen, die eine Person zeigen, sehr zuverlässig erkannt, bei elf Sequenzen ist die Klassifikation nur unter Vorbehalt möglich.

Bis auf die beiden Sequenzen *PKW-6* und *PKW-7* werden alle PKW-Sequenzen zuverlässig erkannt. Diese zeigen einen Lieferwagen, für den nur sehr wenige ähnliche Objekte in der Datenbank gespeichert sind. Bei einem großen Anteil der Sequenzen, die deformierbare Objekte wie beispielsweise Vögel oder Personen zeigen, ist das Maß für die Zuverlässigkeit deutlich geringer. Beide analysierten Vogelsequenzen können trotz manueller – und somit perfekter Segmentierung – nicht erkannt werden, da sich die in der Datenbank gespeicherten Vogelbilder signifikant von den meisten Bildern der Sequenzen unterscheiden. Die Kontur einer Person ist im Vergleich zur Kontur eines PKWs deutlich komplexer, so dass mit dem ursprünglichen Ansatz trotz umfangreicher Datenbank bei einem Vergleich der Skalenraumabbildungen nur die Hälfte der Sequenzen zuverlässig klassifiziert werden können. Beispiele für korrekt klassifizierte Objekte der Sequenzen *PKW-4*, *Person-1* und *Person-4* sind in Abbildung 5.11 dar-

gestellt.

Bei einer Aggregation der Klassifikationsergebnisse über alle Testsequenzen wird deutlich, dass von den 1878 gültigen Bildern nur 69 % korrekt klassifiziert werden, d. h. fast ein Drittel aller Bilder wird fehlerhaft klassifiziert. Der durchschnittliche Wert für die Zuverlässigkeit aller Sequenzen liegt mit 0,64 nur geringfügig über der gewünschten Grenze von 0,6. In den folgenden Abschnitten werden Ergebnisse für die neuen verbesserten Verfahren vorgestellt, durch die eine deutliche Verringerung der Fehler erreicht wird.

5.9.4 Erweiterung des Skalenraumvergleichs durch zusätzliche Merkmale

Durch die zusätzliche Betrachtung der globalen Konturdeskriptoren und eines weiteren Merkmalswertes für jeden Bogen kann eine Verbesserungen der Klassifikationsergebnisse erreicht werden. Der Einsatz globaler Konturdeskriptoren ermöglicht ein effizientes Ausfiltern von deutlich unterschiedlichen Konturen in einem ersten Schritt. Beim Vergleich der Skalenraumabbildungen wird jeder Bogen um den dritten Merkmalswert *Breite des Bogens* erweitert, so dass stark und schwach gekrümmte konkave Bereiche einer Kontur unterschieden werden können.

Da sowohl die Berechnung als auch der Vergleich der globalen Konturdeskriptoren *Kompaktheit* und *Exzentrizität* nur einen sehr geringen Rechenaufwand erfordert, wird bei Verwendung dieser Maße der durchschnittliche gesamte Rechenaufwand des Erkennungsalgorithmus reduziert. Lediglich bei einer Ähnlichkeit der Deskriptoren werden die komplexeren Vergleiche der Skalenraumabbildungen durchgeführt. Zwei Konturen gelten als ähnlich, falls folgende Bedingungen erfüllt sind:

$$\frac{\max(C_{OB(i)}, C_{DB(j)})}{\min(C_{OB(i)}, C_{DB(j)})} < T_C \quad \text{und} \quad (5.21)$$

$$\frac{\max(E_{OB(i)}, E_{DB(j)})}{\min(E_{OB(i)}, E_{DB(j)})} < T_E. \quad (5.22)$$

Bei dem Vergleich eines Objektes i mit einem Element j der Datenbank dürfen die Werte für die Kompaktheit C und die Exzentrizität E nicht allzu deutlich voneinander abweichen. Die Faktoren, um die sich beide Werte maximal unterscheiden dürfen, liegen bei $T_C = 1,5$ bzw. $T_E = 1,3$. Mit den gewählten Schwellwerten werden durchschnittlich 85 Prozent der Bilder der

Sequenz:
PKW-4



ähnlichstes
Objekt der
Datenbank



Sequenz:
Person-1



ähnlichstes
Objekt der
Datenbank



Sequenz:
Person-4



ähnlichstes
Objekt der
Datenbank



Abbildung 5.11: Ausgewählte Klassifikationsergebnisse der Testsequenzen PKW-4 (oben), Person-1 (Mitte) und Person-4 (unten). Für jedes segmentierte Objekt der Videosequenz wird das ähnlichste Objekt der Datenbank angezeigt.

Sequenz		Durchschnittliche Anzahl der Objekte der Datenbank nach der Filterung mit den globalen Deskriptoren					
		Säugetier	Vogel	Flugzeug	Schiff	PKW	Person
1–6	Person	4	3	3	6	10	25
7–8	Vogel	8	6	4	5	14	13
9–15	PKW	1	1	1	6	23	11
16–26	Person	6	4	4	6	12	26
27–30	PKW	1	1	1	6	26	9

Tabelle 5.3: Durchschnittliche Anzahl der Objekte der Datenbank nach der Filterung mit den globalen Konturdeskriptoren

Datenbank verworfen, so dass in diesen Fällen der Vergleich der Bögen der Skalenraumabbildungen nicht durchgeführt wird. Tabelle 5.3 gibt an, wieviele Bilder der Datenbank nach dem Vergleich mit den Konturdeskriptoren bei den unterschiedlichen Sequenzen durchschnittlich pro Bild erhalten bleiben. Durch den Vergleich der Konturdeskriptoren werden viele deutlich unterschiedliche Konturen zuverlässig und schnell ausgefiltert. Eine Klassifikation ist durch die hohe Varianz der Konturdeskriptoren innerhalb einer Objektklasse jedoch nicht möglich.

Für ähnliche Konturdeskriptoren wird ein Vergleich der Skalenraumabbildungen durchgeführt. Jeder Bogen einer Skalenraumabbildung wird durch die drei Werte *Position*, *Höhe* und *Breite* charakterisiert. Nur wenn alle Parameter ähnlich sind, ist der Vergleich zweier Bögen erfolgreich, und ein Differenzwert wird berechnet. In der linken Hälfte von Tabelle 5.4 sind die Klassifikationsergebnisse unter Berücksichtigung der globalen Konturdeskriptoren und der Breite der Bögen der Skalenraumabbildungen angegeben. Der Anteil der gültigen Bilder sinkt von durchschnittlich 99 Prozent beim einfachen Skalenraumvergleich auf 96 Prozent. Insbesondere stark fehlerhaft segmentierte Objekte und solche, für die keine ähnlichen Objekte in der Datenbank enthalten sind, werden in diesem Schritt entfernt.

Der Anteil der korrekt klassifizierten Objekte steigt von durchschnittlich 69 Prozent auf 75 Prozent. Bei den PKW-Sequenzen liegt der Anteil der korrekt klassifizierten Objekte sogar über 90 Prozent. In zwei Sequenzen sinkt der Anteil der korrekt erkannten Objekte geringfügig, da durch die zusätzlichen Konturmerkmale bei ungenau segmentierten Objekten auch korrekte Klassifikationsergebnisse verworfen werden können. Bezogen auf alle Sequenzen ist die Auswirkung jedoch sehr gering, so dass der Anteil der korrekt klassifizierten Objekte durchschnittlich um 6 Prozent steigt.

In das Maß für die Zuverlässigkeit entsprechend der Gleichung 5.19 geht der Unterschied zwischen der am häufigsten und der am zweithäufigsten erkannten Objektklasse ein. Die An-

Nr.	Klassifikationsergebnisse mit globalen Konturdeskriptoren			Klassifikationsergebnisse mit transformierten Konturen		
	Anteil gültiger Bilder	erkannte Objekt-klasse	Maß für die Zuverlässigkeit	Anteil gültiger Bilder	erkannte Objekt-klasse	Maß für die Zuverlässigkeit
1	100%	100%	1,00	100%	100%	1,00
2	95%	97%	0,95	97%	95%	0,89
3	97%	66%	0,52 (+0,07)	74%	69% (+7%)	0,60 (+0,15)
4	100%	79%	0,70 (-0,06)	93%	93% (+17%)	0,92 (+0,16)
5	92%	100% (+31%)	1,00 (+0,20)	69%	100% (+31%)	1,00 (+0,20)
6	99%	83% (+24%)	0,80 (+0,27)	88%	81% (+22%)	0,83 (+0,30)
7	100%	33%	0,11 (+0,11)	67%	60% (+27%)	0,50 (+0,50)
8	75%	58% (-4%)	0,16 (-0,08)	49%	52% (-10%)	0,10 (-0,14)
9	100%	100%	1,00	97%	100%	1,00
10	100%	100%	1,00	100%	100%	1,00
11	100%	100%	1,00	100%	100%	1,00
12	100%	100%	1,00	84%	100%	1,00
13	95%	100%	1,00	91%	100%	1,00
14	58%	67% (+12%)	0,33	18%	100% (+45%)	1,00 (+0,65)
15	79%	91% (+14%)	0,82 (+0,28)	21%	100% (+23%)	1,00 (+0,46)
16	82%	69%	0,52	31%	75% (+11%)	0,64 (+0,17)
17	98%	88%	0,80 (-0,05)	90%	95% (+7%)	0,95 (+0,10)
18	99%	85% (+9%)	0,71	93%	95% (+19%)	0,95 (+0,22)
19	82%	74% (+10%)	0,70 (+0,06)	54%	80% (+16%)	0,85 (+0,21)
20	99%	56% (+12%)	0,29 (+0,18)	89%	70% (+26%)	0,52 (+0,41)
21	99%	48% (-6%)	0,53 (+0,08)	62%	63% (+9%)	0,71 (+0,26)
22	100%	90% (+51%)	0,87 (+0,46)	65%	85% (+46%)	0,89 (+0,48)
23	97%	65%	0,52	97%	59%	0,60
24	100%	69%	0,70	89%	71%	0,81 (+0,08)
25	99%	75%	0,67	90%	84% (+14%)	0,85 (+0,20)
26	89%	52% (+9%)	0,37	79%	59% (+16%)	0,53 (+0,20)
27	100%	92%	0,83	92%	100% (+8%)	1,00 (+0,17)
28	93%	100% (+14%)	1,00 (+0,29)	79%	100% (+14%)	1,00 (+0,29)
29	100%	100%	1,00	100%	100%	1,00
30	90%	96%	0,93 (+0,07)	83%	96%	0,92 (+0,06)
Σ	96%	75% (+6%)	0,69 (+0,05)	81%	82% (+13%)	0,83 (+0,19)

Tabelle 5.4: Klassifikationsergebnisse zur Objekterkennung mit Optimierungen. Fehlerhafte oder unzuverlässige Klassifikationsergebnisse sind fett markiert.

zahl der zuverlässig und korrekt erkannten Sequenzen steigt von 18 auf 21, und nur noch die Sequenz *Vogel-2* wird fehlerhaft klassifiziert. Durchschnittlich steigt das Maß für die Zuverlässigkeit von 0,64 auf 0,69.

Obwohl der Anteil der korrekt erkannten Objekte deutlich erhöht wurde, werden immer noch 25 Prozent fehlerhaft klassifiziert. Da die Skalenraumabbildungen lediglich Merkmale für die konkaven Bereiche einer Kontur liefern, fehlen bei vielen Konturen wichtige Informationen für eine zuverlässige Beschreibung dieser.

5.9.5 Klassifikation mit transformierten Konturen

Die Ergebnisse der Objekterkennung mit transformierten Konturen werden in diesem Abschnitt analysiert. Da bei der Analyse einer transformierten Kontur Merkmale für konkave und konvexe Objektregionen berücksichtigt werden, sind deutlich zuverlässigere Klassifikationen möglich. Nach dem Vergleich der globalen Konturdeskriptoren werden die Bögen zweier Skalenraumabbildungen der ursprünglichen und der transformierten Kontur verglichen. Durch die Transformation sind durchschnittlich doppelt so viele Merkmalswerte zur Beschreibung der Kontur verfügbar. Viele Objekte, die beim Vergleich mit dem ursprünglichen Verfahren große Ähnlichkeiten besaßen, weisen jetzt deutliche Unterschiede auf. Dadurch steigt trotz identischer Schwellwerte der Anteil der ungültigen Bilder signifikant. Detaillierte Ergebnisse sind in der rechten Hälfte von Tabelle 5.4 ersichtlich.

In zehn Sequenzen sinkt die Anzahl der gültigen Bilder unter 75 Prozent. Mehrere Ursachen sind für den hohen Anteil verantwortlich: Eine *fehlerhafte Segmentierung* tritt insbesondere in den automatisch segmentierten Sequenzen auf, bei denen Teile des Objektes und des Hintergrundes ähnliche Helligkeitswerte annehmen. Zusätzlich können die Fehler durch den *Schatten des Objektes* verstärkt werden. *Fehlende Details* einer Kontur treten insbesondere bei Objekten mit geringer Größe auf. Eine teilweise *Verdeckung eines Objektes*, z. B. durch eine Straßenlaterne oder ein Schild, verursacht Segmentierungsfehler in einzelnen Bildern. Es ist möglich, dass ein Objekt erst im Bild *erscheint* bzw. dieses *verlässt* und in einzelnen Bildern nicht vollständig sichtbar ist. Korrekt segmentierte Objekte können nicht erkannt werden, wenn *keine ähnlichen Objekte in der Datenbank* gespeichert sind. Dies gilt insbesondere für viele Bilder der Sequenzen 7 und 8 bzw. 14 und 15, die einen Vogel bzw. einen Lieferwagen zeigen. Abbildung 5.12 verdeutlicht Beispiele ungültiger Objekte einzelner Sequenzen, für die kein ähnliches Objekt in der Datenbank gefunden wurde.

Durchschnittlich steigt der Anteil der korrekt erkannten Objekte von 69 auf 82 Prozent. Se-

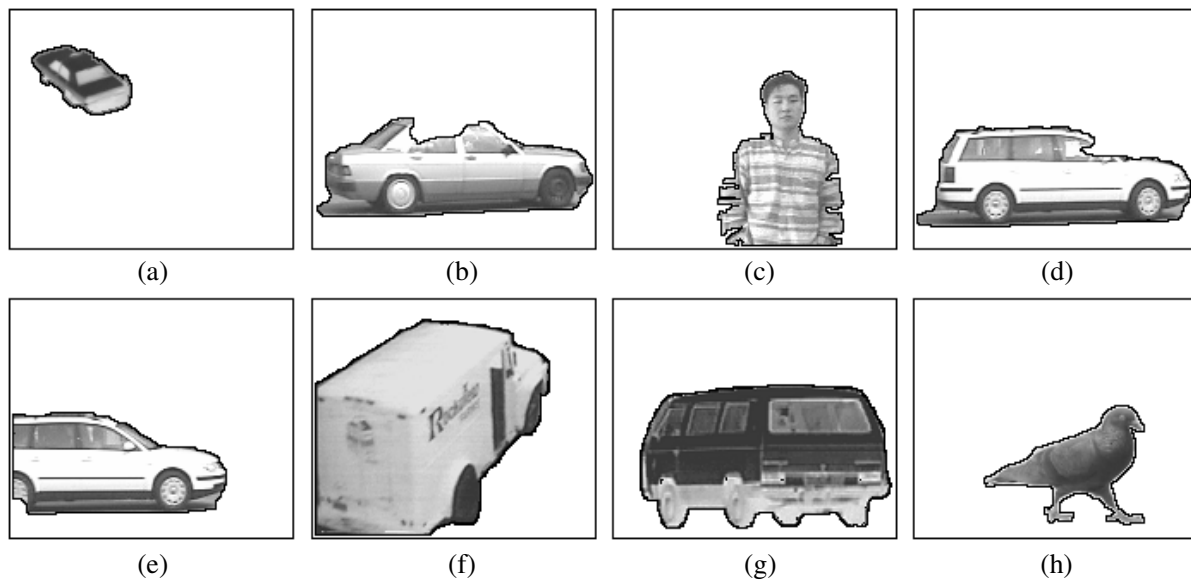


Abbildung 5.12: Beispiele ungültiger Objekte, für die kein ähnliches Objekt in der Datenbank gefunden wurde. Typische Fehler resultieren aus fehlenden Details bei Objekten mit geringer Größe (a), Segmentierungsfehlern (b,c), Schatten (b,d,e), nur teilweise sichtbaren Objekten (e,f) oder aus fehlenden ähnlichen Objekten in der Datenbank (f,g,h).

quenz 8 wird weiterhin fehlerhaft klassifiziert, wobei das Maß für die Zuverlässigkeit nur knapp über null liegt und das Ergebnis nicht aussagekräftig ist. Auch der durchschnittliche Wert der Zuverlässigkeit steigt bei der Klassifikation mit transformierten Konturen um 0,19 auf 0,83. Bis auf die beiden Vogelsequenzen und zwei Sequenzen mit Personen können alle Testsequenzen sehr zuverlässig erkannt werden.

5.9.6 Objekterkennung in historischen Videos

Die Algorithmen zur Segmentierung und Erkennung von Objekten wurden in das *European-Chronicles-Online*-System integriert, das in Kapitel 2.3.6 vorgestellt wurde. Beim Einfügen eines neuen Videos in das Archiv wird die Objekterkennung automatisch gestartet und die Informationen über erkannte Objekte nach Abschluss der Berechnung im System gespeichert. Der wesentliche Vorteil beim Einsatz automatischer Algorithmen zur Objekterkennung liegt darin, dass Informationen über Videos ohne zusätzlichen Aufwand für die Archivare zur Verfügung gestellt werden. Im Rahmen des *European-Chronicles-Online*-Projektes wurden mehr als 1200 historische Videos analysiert und die Objektinformationen im Archiv gespeichert. Beispiele für korrekt klassifizierte Objekte der historischen Videos sind in Abbildung 5.13

dargestellt.

Historische Videos stellen eine besondere Herausforderung für Algorithmen zur Segmentierung und Klassifikation von Objekten dar. *Streifen* und *Kratzer* sowie starkes *Rauschen* führen in einzelnen Bildern zu deutlichen Bildfehlern, so dass eine zuverlässige Schätzung der Parameter des Kameramodells nicht immer möglich ist. Bilder mit fehlerhaften Kameraparametern werden mit Hilfe der in Kapitel 3.5 vorgestellten Algorithmen zuverlässig identifiziert und ausgefiltert.

Durch die zum Teil sehr geringe Bildqualität historischer Videos ist auch bei einem korrekten Kameramodell die präzise Segmentierung der Objekte eine große Herausforderung. Insbesondere bei einem geringen *Kontrast*, bei *Bildfehlern*, bei *Helligkeitsschwankungen* und bei *unscharfen Aufnahmen* sind die Objektgrenzen schwer zu identifizieren, so dass durch den Vergleich mit dem Hintergrundbild ungenau segmentierte Objekte entstehen.

Da die Objekterkennungsalgorithmen die Informationen über die Objekte automatisch ermitteln und ohne Benutzerinteraktion im *European-Chronicles-Online*-Archiv speichern, sollte der Anteil der korrekt klassifizierten Objekte (Präzision) möglichst hoch sein. Der Wert für die Vollständigkeit des Algorithmus ist von geringerer Bedeutung, da Suchanfragen häufig mehr als einhundert passende Videosequenzen finden, von denen wegen des erforderlichen Zeitaufwands im Allgemeinen nur einzelne tatsächlich betrachtet werden.

Um den Anteil der fehlerhaft klassifizierten Objekte gering zu halten, werden niedrige Schwellwerte in der Gleichung 5.13 für die maximal zulässigen Unterschiede bezüglich der *Höhe*, der *Position* und der *Breite* der Bögen der Skalenraumabbildungen angesetzt ($T_P = T_H = 15\%$, $T_W = 30\%$), so dass mit hoher Wahrscheinlichkeit korrekt klassifizierte Objekte in das *European-Chronicles-Online*-System übernommen werden.

Bezogen auf Kameraeinstellungen wird durch die gewählten Schwellwerte eine *Präzision* für die Erkennung von PKWs und Personen in den 1200 analysierten historischen Videos von über 96 Prozent erreicht. Das Maß für die *Vollständigkeit* bei PKWs und Personen liegt wegen der großen Anzahl fehlerhafter Kameraparameter und der ungenauen Segmentierung bei 21 Prozent. Für Flugzeuge und Schiffe sind die Werte für die Präzision und Vollständigkeit deutlich niedriger, da der Himmel bzw. die Wellen nur wenige Merkmalspunkte zur Berechnung der Kameraparameter liefern. Um den Anteil der fehlerhaften Daten im Archiv gering zu halten, wird die automatische Objekterkennung im *European-Chronicles-Online*-System standardmäßig nur für PKWs und Personen eingesetzt.

Trotz des relativ geringen Wertes für die Vollständigkeit sind die Algorithmen zur Objekterkennung eine sehr große Hilfe bei der Indexierung von Videoarchiven. Jede zusätzliche Infor-

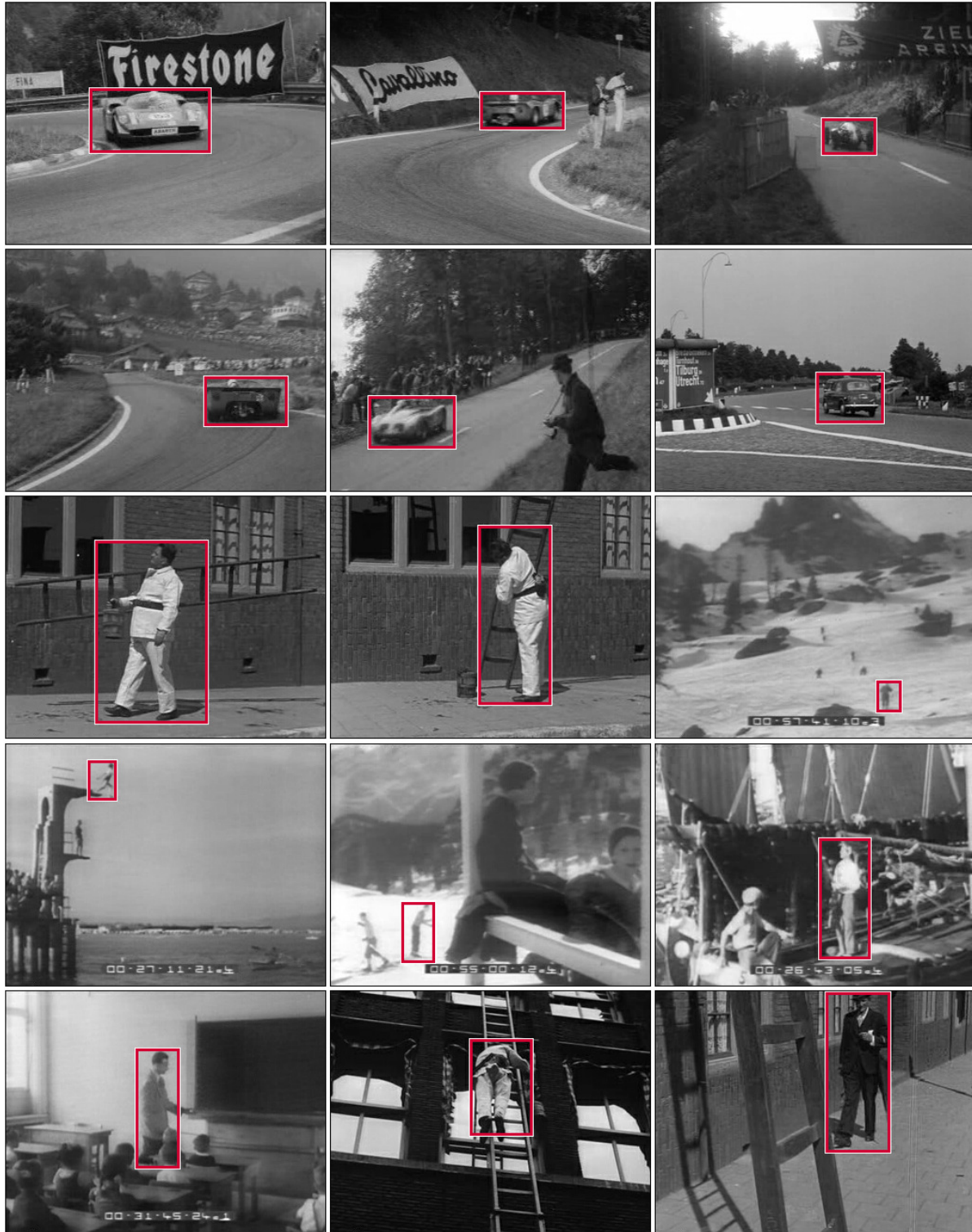


Abbildung 5.13: Beispiele für automatisch segmentierte und klassifizierte Objekte des European-Chronicles-Online-Videoarchivs.

mation über ein Video verbessert die Zugriffs- und Suchmöglichkeiten, wodurch insbesondere die Arbeit mit umfangreichen Videoarchiven erleichtert wird. Die Präzision der Algorithmen zur Objekterkennung ist sehr hoch, so dass bei Suchanfragen nur vereinzelt fehlerhafte Ergebnisse angezeigt werden.

5.10 Zusammenfassung

In diesem Kapitel wurden Verfahren zur Klassifikation von Objekten in Videos vorgestellt. Dazu wurden charakteristische Merkmale mit Hilfe des Skalenraumansatzes anhand der äußeren Kontur eines Objektes ermittelt. Ein wesentlicher Vorteil dieses Verfahrens besteht darin, dass es die menschliche Wahrnehmung bei der Beurteilung der Ähnlichkeiten zweier Konturen sehr gut annähert.

Wir haben zwei neue Verfahren entwickelt, um wesentliche Probleme des Skalenraumansatzes zu beheben: Um zu verhindern, dass unterschiedlich stark konkav gekrümmte Bereiche von Konturen zu identischen Merkmalswerten führen, wird im ersten Verfahren die Breite der Bögen in den Skalenraumabbildungen als neues Merkmal eingeführt. Das zweite neue Verfahren berechnet transformierte Konturen und leitet Merkmale zur Beschreibung konvexer Objektregionen ab. Dadurch wird sogar eine Erkennung konvexer Objekte möglich. Zur Klassifikation von Videosequenzen haben wir ein neues Verfahren zur Aggregation der Ergebnisse der Einzelbilder entwickelt, bei dem die Distanzen zwischen einem unbekannten Objekt und den Objektklassen der Datenbank berechnet und aggregiert werden.

Im Rahmen der experimentellen Ergebnisse wurde anhand von 30 Videosequenzen ein Vergleich des ursprünglichen Skalenraumansatzes und der neuen Verfahren durchgeführt, wobei der Anteil der korrekt erkannten Einzelbilder in den Videosequenzen von 69 Prozent auf über 82 Prozent steigt. Nach der Aggregation der Einzelergebnisse erhöht sich der Anteil der korrekt und zuverlässig erkannten Videosequenzen von 60 Prozent beim ursprünglichen Skalenraumansatz auf über 86 Prozent bei unserem neu entwickelten Verfahren. Zusätzlich wurden die Algorithmen zur Objekterkennung in das *European-Chronicles-Online*-System integriert und ermöglichen eine zuverlässige automatische Erkennung von Personen und PKWs in den historischen Videos des Archivs.

Zwei neue Anwendungen, welche die Objekterkennungsalgorithmen nutzen, werden in Kapitel 10 vorgestellt. Dabei werden detaillierte Bewegungen einer Person und die Fahrt eines PKWs automatisch analysiert [279]. Auch bei diesen Anwendungen erfolgt die Klassifikation mit transformierten Konturen, da sie deutlich zuverlässigere und genauere Ergebnisse liefern.

KAPITEL 6

Erkennung von Textregionen und Buchstaben

Schon seit vielen Jahren existieren Softwareprodukte, um Texte in hochauflösenden eingescannten Dokumenten automatisch zu erkennen. Die Erkennung von Buchstaben (*OCR*, engl. *optical character recognition*) funktioniert für Textseiten mit monochromem Hintergrund sehr zuverlässig. Andere Systeme wurden erfolgreich zur Erkennung von Nummern- oder Straßenschildern entwickelt. Diese sehr spezialisierten Verfahren sind im Allgemeinen nicht geeignet, Texte in Videos oder in Bildern mit komplexem Hintergrund zu erkennen. Dabei liefern Texte besonders wichtige semantische Informationen über ein Video. Beispielsweise nennen Texte in Nachrichtensendungen den Namen von Orten oder Personen und eignen sich daher besonders gut zur Indexierung eines Videos.

Das Ziel dieses Kapitels soll es nicht sein, eine Texterkennungssoftware mit vergleichbarer Genauigkeit wie aktuelle OCR-Systeme bei der Erkennung eingescannter Dokumente zu entwickeln. Dieses Vorhaben wäre vom Umfang her nicht innerhalb dieser Arbeit zu realisieren. Vielmehr werden einzelne interessante Fragestellungen detailliert analysiert und neue Ideen für ausgewählte Teilprobleme entwickelt.

Besondere Probleme entstehen bei der Erkennung von Texten durch die *geringe Auflösung* der Bilder. Im Vergleich zu eingescannten Bildern stehen deutlich weniger Pixel zur Beschreibung eines Buchstabens zur Verfügung. Ein weiteres Problem sind Bildfehler und unscharfe Kanten durch hohe *Kompressionsraten*. Das Ausfiltern hoher Frequenzen verwischt die Buchstaben

mit dem Hintergrund und mit benachbarten Buchstaben. Im Gegensatz zu eingescannten Dokumenten enthält der Hintergrund in Bildern und Videos häufig *komplexe Texturen*, die eine exakte Segmentierung der Buchstaben erschweren. Die *Größe* und der *Zeichensatz* der Texte kann in Bildern und Videos deutlich variieren. Auch *Rauschen*, d. h. das Auftreten von einzelnen zufällig verteilten Pixelfehlern, ist in Digitalfotos und digitalen Filmen deutlich stärker ausgeprägt als bei Scannern. Rauschen wird durch *ungünstige Lichtverhältnisse* und die schlechte Ausleuchtung bei der Aufnahme zusätzlich verstärkt. Des Weiteren liegt der Fokus eines Bildes nicht immer innerhalb einer Textregion, so dass Bereiche mit Texten eine *geringe Bildschärfe* aufweisen können. Falls ein Text innerhalb eines Bildes nicht parallel zur Bildebene liegt, erhöhen *affine* und *perspektivische Verzerrungen* des Textes die Komplexität der Erkennung.

Zwei Arten von Texten werden in Bildern oder Videos unterschieden, *Texte innerhalb von Szenen* (engl. *scene text*) – wie Straßen- bzw. Gebäudeschilder oder die Schrift auf einem T-Shirt – oder *künstlich überlagerte Texte* (engl. *graphic text* oder *superimposed text*). Überlagerte Texte stellen häufig zusätzliche semantische Informationen zur Verfügung, die im Video nicht enthalten sind. In einer Nachrichtensendung sind die Namen von Politikern oder Orten typische Beispiele für überlagerte Texte. Die besondere Schwierigkeit bei der Erkennung eines Szenentextes liegt darin, dass der Text nicht senkrecht zur Kamera ausgerichtet ist, sondern in alle drei Dimensionen gekippt sein kann [87, 88, 372].

Im folgenden Abschnitt werden zunächst Verfahren zum Auffinden von Textregionen vorgestellt. Abschnitt 6.2 beschreibt die Erkennung von Textregionen mit Hilfe von Projektionsprofilen. Eine besondere Herausforderung in Bildern oder Videos mit komplexem Hintergrund ist die korrekte Segmentierung eines einzelnen Buchstabens. In Abschnitt 6.3 werden zwei neue Algorithmen zur Verbesserung der Segmentierung vorgestellt. Durch einen optimierten Kürzeste-Pfade-Algorithmus werden zunächst Trenner zwischen einzelnen Buchstaben identifiziert. Zur Unterscheidung zwischen einem Text- und Hintergrundpixel wird ein modifizierter Region-Merging-Algorithmus eingeführt, der als Distanzmaß ähnliche Farben und die Entfernung zwischen Bildregionen berücksichtigt. Vier Verfahren zur Klassifikation von Buchstaben werden in Abschnitt 6.4 vorgestellt und im Rahmen der experimentellen Ergebnisse analysiert. Dabei liefern Skalenraumabbildungen mit transformierten Konturen besonders zuverlässige Ergebnisse.

6.1 Existierende Verfahren zur Texterkennung

Bei der Erkennung von Textregionen werden zwei wesentliche Ansätze unterschieden, die Analyse von *Texturen* und die Aggregation ähnlicher *Regionen*. Bei den Verfahren der ersten Gruppe werden starke *Kanten*, *Ecken* oder Pixel mit einem hohen *Kontrast* ermittelt [66, 158, 416]. Auch eine Analyse komprimierter Bilddaten – insbesondere die hochfrequenten DCT-Koeffizienten – ermöglicht die Erkennung von Textregionen [99, 481, 583, 589]. Die Verwendung von Textmerkmalen hat den Nachteil, dass eine große Anzahl von fehlerhaft erkannten Textregionen in Bildern mit komplexem Hintergrund auftritt. Bei der Aggregation ähnlicher Textregionen werden Bildbereiche mit ähnlichen Farben gesucht und Textpixel anhand spezieller Heuristiken (Buchstabengröße, Mindestkontrast, räumliche Anordnung einzelner Zeichen) ermittelt [320, 567].

Durch die Analyse mehrerer Bilder im Zeitablauf wird eine zuverlässigere Erkennung von Textregionen in Videos möglich [163]. Ein Text ist immer in mehreren hintereinander folgenden Bildern sichtbar, da er sonst nicht gelesen werden könnte. Die Bewegung der Texte ist auf horizontale oder vertikale Verschiebungen beschränkt. Es stehen viele Techniken zur Verfügung, um segmentierte Buchstaben in Graustufen- oder Binärbildern zu erkennen [116, 173, 375, 505]. Bekannte Verfahren wie die Fourier-, DCT- oder Wavelet-Transformationen, die Karhunen-Loève-Transformation oder Konturprofile sind zur Klassifikation von Buchstaben geeignet. Mehrere umfangreiche Publikationen über die unterschiedlichen Verfahren zur Erkennung von Buchstaben wurden in den letzten Jahren veröffentlicht [126, 319, 345, 376].

Hua et al. haben eine Kombination mehrerer Verfahren zur Erkennung von Textregionen eingesetzt [214, 218]. Zunächst werden in einem texturbasierten Ansatz starke Ecken im Bild ermittelt und mit benachbarten Ecken zu möglichen Textregionen zusammengefasst. Zur Erkennung der Textregionen in Videos werden nur einzelne Bilder ausgewählt, die eine besonders gute Segmentierung erwarten lassen und einen hohen Kontrast innerhalb der Textregionen enthalten. Zusätzlich führen die Autoren noch ein Maß zur Beurteilung der Genauigkeit der Segmentierung ein [217].

Mehrere Algorithmen zur Erkennung von Textregionen wurden von Lienhart et al. entwickelt. Mit Hilfe eines regelbasierten Ansatzes werden mögliche Textregionen anhand ihres Kontrastes, der Textfarbe und der Buchstabengröße bestimmt [320]. In einem weiteren Verfahren wird ein mehrstufiges neuronales Netz trainiert, das ein Bild in unterschiedlichen Skalierungen analysiert und Textregionen erkennt [324, 543]. Neuronale Netze, die als Eingabe Wavelet- oder DCT-Koeffizienten verwenden, werden auch in mehreren anderen Erkennungsalgorithm-

men eingesetzt [307, 383, 560].

Insbesondere für eingeschränkte Anwendungsszenarien gibt es erfolgreiche Systeme zur automatischen Segmentierung und Erkennung von Texten. Ein Schwerpunkt liegt in der Analyse von Nachrichtensendungen, da der Anteil textueller Informationen besonders hoch ist und durch die gute Strukturierung der Sendungen die Erkennung erleichtert wird. Xi et al. verwenden Kantenbilder und morphologische Operatoren, um Textregionen zu identifizieren [552]. Sato et al. verbessern innerhalb eines Videos zunächst die Bildqualität einzelner vergrößerter Bilder durch Subpixel-Interpolation und Aggregation über mehrere Bilder [450]. Vier Filter liefern hierbei eine Schätzung für die Positionen der Textregionen, wobei die genauen Grenzen der Buchstaben durch Projektionsprofile ermittelt werden. Bei den Ansätzen von Antani et al. werden mehrere Verfahren kombiniert, unter anderem 4x4-Blockfilter, DCT-Koeffizienten und Algorithmen zur Aggregation zusammenhängender Regionen [13, 14, 159].

Das *Text-Finder*-System analysiert die Textur von Regionen und aggregiert ähnliche Regionen mit dem K-Means-Algorithmus [531]. Buchstaben werden durch besonders stark ausgeprägte Kanten identifiziert, die zu Regionen zusammengefasst werden. Weitere spezialisierte Anwendungen ermöglichen die Erkennung von Straßenschildern und Firmennamen [157, 563, 551], Nummernschildern von Fahrzeugen [102], ausgefallenen Schriftarten [338, 480] oder mathematischen Zeichen [489].

In den meisten Erkennungssystemen wird angenommen, dass ein monochromer Hintergrund vorliegt, so dass die Segmentierung eines einzelnen Buchstabens sehr zuverlässig funktioniert. Insbesondere in natürlichen Bildern oder – mit Ausnahme von Nachrichtensendungen – Videos trifft diese Annahme jedoch nur selten zu. In fast allen vorgestellten Ansätzen bleibt der letzte Schritt – die Erkennung der einzelnen Buchstaben – unberücksichtigt, und in den meisten Veröffentlichungen wird lediglich auf kommerzielle OCR-Systeme verwiesen. Um nicht auf externe OCR-Systeme angewiesen zu sein, erfolgt die Erkennung einzelner Buchstaben mit den von uns entwickelten Verfahren. Bei der eigenständigen Entwicklung einer OCR-Software bleibt kritisch anzumerken, dass kommerzielle OCR-Systeme eine äußerst zuverlässige Texterkennung bei eingescannten Dokumenten ermöglichen, was durch Optimierungen und Verbesserungen über viele Jahre erreicht wurde. Diese hervorragenden Erkennungsraten werden bei der Texterkennung in Bildern und Videos von uns bei weitem nicht erreicht.

6.2 Erkennung von Textregionen

Die Erkennung von Buchstaben in Bildern und Videos erfolgt in drei Schritten, auf die in den folgenden Abschnitten eingegangen wird. Im ersten Schritt, bei dem mögliche *Textregionen* identifiziert und durch rechteckige Regionen beschrieben werden, wird auf bekannte Verfahren zurückgegriffen. Es werden die Annahmen getroffen, dass mehrere Wörter in jeder Textzeile enthalten sind und dass ein starker Kontrast zwischen Buchstaben und Bildhintergrund besteht. Anschließend werden zur Segmentierung der einzelnen Buchstaben *Trenner zwischen den Buchstaben* gesucht, um zu verhindern, dass zwei oder mehrere Buchstaben zusammenhängende Regionen bilden. Ein modifizierter *Region-Merging*-Algorithmus klassifiziert die einzelnen Pixel als Text oder Hintergrund. Im letzten Schritt wird die äußere Kontur eines Buchstabens analysiert und mit Hilfe eines Skalenraumvergleiches klassifiziert.

Um eine Textregion zu erkennen, wird die von Sato und Smith vorgestellte Technik verwendet, bei der Textregionen anhand ihres hohen Kontrastes und ihrer starken Kanten gesucht werden [449, 464]. Es wird die Annahme getroffen, dass jede Textzeile mehrere Wörter enthält und ein deutlicher Kontrast zwischen Text und Hintergrund besteht. Ein Filter läuft horizontal über das Bild und markiert Blöcke mit starken Kanten. Zusammenhängende Blöcke definieren Textregionen, die bestimmte Kriterien in Bezug auf ihre Größe erfüllen müssen. Jede Textregion wird durch ein umgebendes Rechteck beschrieben.

Es ist möglich, dass mehrere Textzeilen in einer Textregion enthalten sind, so dass in einem zweiten Schritt die exakte obere und untere Grenze einer Textzeile mit Hilfe von *Projektionsprofilen* (engl. *projection profile*) bestimmt wird [324, 450]. Ein Projektionsprofil ist definiert als Summe der absoluten Differenzwerte benachbarter Pixel. Durch die Übergänge zwischen Text und Bildhintergrund sind innerhalb einer Textzeile die Differenzen benachbarter horizontaler Pixel besonders hoch. Hohe Beträge geben einen Hinweis auf eine Textzeile, bei sehr niedrigen Werten kann ein Bereich ohne Text angenommen werden. Besonders deutliche Änderungen der Profilwerte treten am oberen und unteren Rand einer Textzeile auf. Abbildung 6.1 zeigt ein horizontales Projektionsprofil mit stark ausgeprägten Profilwerten im Bereich der Textzeile.

Die Ergebnisse der einzelnen Schritte bei der Erkennung von Textregionen sind in Abbildung 6.2 dargestellt. Neben den Textregionen sind weitere stark texturierte Bildbereiche ausgewählt. Die Analyse der Projektionsprofile entfernt diese Hintergrundbereiche und ermöglicht eine zuverlässige Identifikation der einzelnen Textzeilen.



Abbildung 6.1: Horizontales Projektionsprofil zur Erkennung einer Textzeile

6.3 Segmentierung von Buchstaben

Eine genaue Segmentierung der einzelnen Buchstaben innerhalb einer Textzeile ist für eine gute Klassifikation besonders wichtig. Schon bei geringen Fehlern ergeben sich deutliche Unterschiede in der Kontur, so dass eine Erkennung nicht mehr möglich ist. Zwei neue Algorithmen zur Segmentierung von Buchstaben werden im Folgenden vorgeschlagen. Zunächst werden, wie schon erwähnt, Trenner zwischen Buchstaben festgelegt, um zu verhindern, dass zwei oder mehrere Buchstaben eine zusammenhängende Einheit bilden. Anschließend wird die dominante Textfarbe durch eine Analyse von Histogrammen bestimmt, und mit Hilfe eines modifizierten Region-Merging-Algorithmus werden die Pixel dann als Text oder Hintergrund klassifiziert.

6.3.1 Ermittlung der Trenner zwischen Buchstaben

Im ersten Schritt der Segmentierung eines Buchstabens werden Trenner identifiziert, welche die Grenzen zwischen benachbarten Buchstaben festlegen. Obwohl vertikale Projektionsprofile zur Erkennung der Buchstabengrenzen in vielen Systemen eingesetzt werden, sind sie für Bilder mit komplexem Hintergrund nicht geeignet. Die Anzahl der fehlerhaft getrennten Buchstaben und der nicht erkannten Trenner ist bei texturiertem Hintergrund sehr groß. Abbildung 6.3 (unten) verdeutlicht typische Fehler bei der Anwendung von Projektionsprofilen.

Der im Folgenden vorgestellte Algorithmus reduziert die Anzahl der nicht erkannten Trenner bzw. der fehlerhaft getrennten Buchstaben signifikant. Im Allgemeinen ist der Kontrast zwischen Text- und Hintergrundpixel sehr hoch, wogegen die Unterschiede innerhalb der Textpixel oder der Hintergrundpixel deutlich geringer sind. Innerhalb der Textzeile wird ein abwärts gerichteter Pfad als Trenner zwischen zwei Buchstaben gesucht. In der obersten Pixelzeile der Textregion werden unterschiedliche Startpositionen für diesen Pfad festgelegt, und für jede Position wird der Pfad zur untersten Pixelzeile mit den jeweils geringsten Kosten berechnet. Die

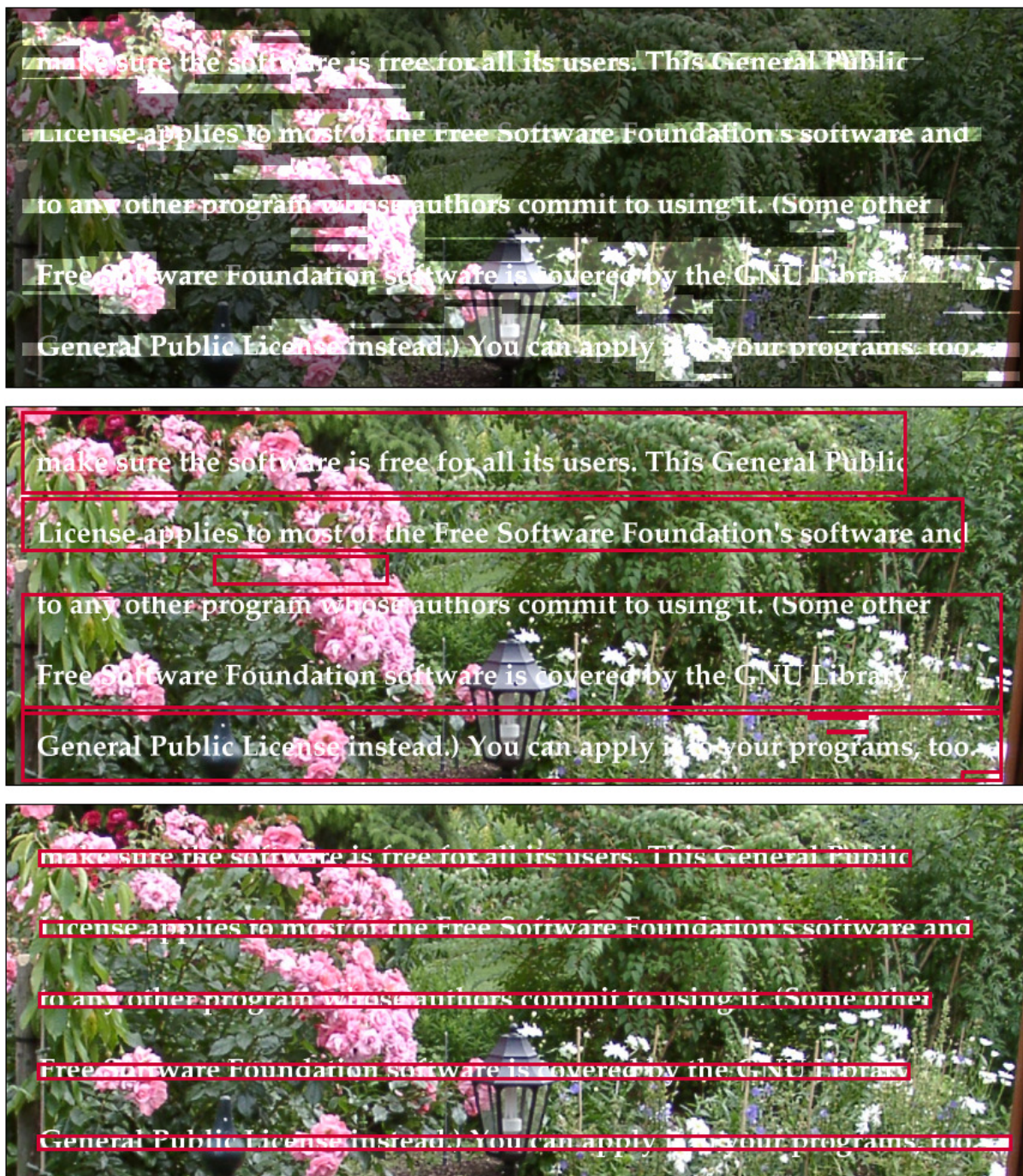


Abbildung 6.2: Erkennung der Textzeilen eines Bildes: Markierung der Blöcke mit starken Kanten (oben), Zuordnung zusammenhängender Blöcke zu Textregionen (Mitte) und Erkennung der Textzeilen durch Projektionsprofile (unten).

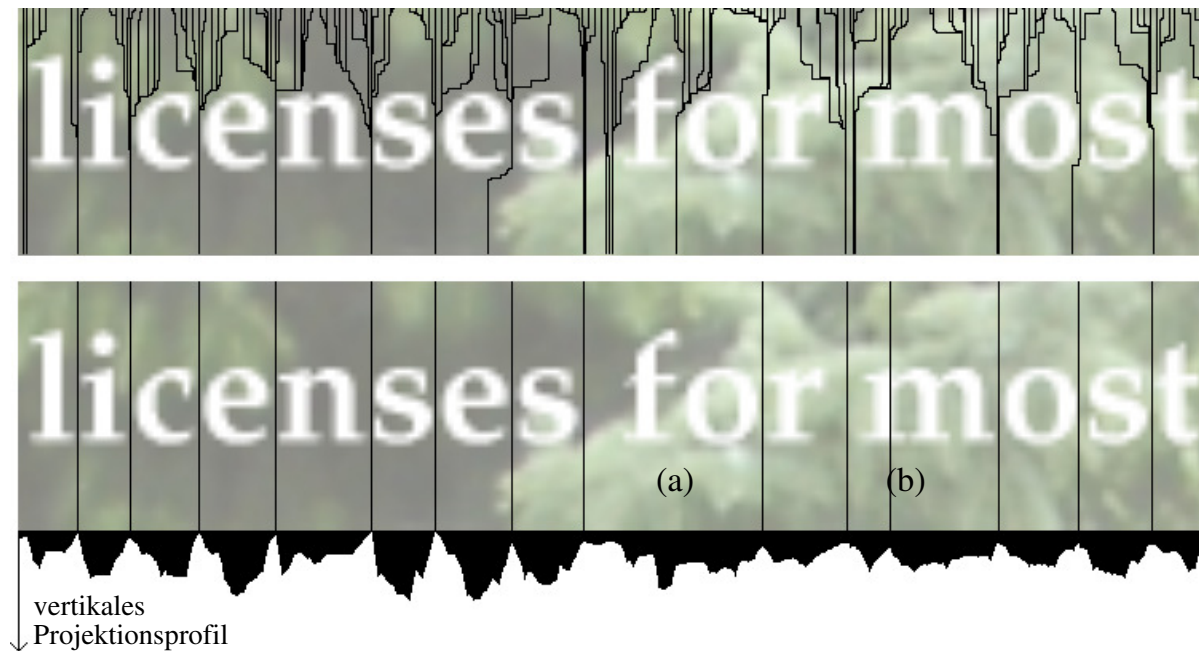


Abbildung 6.3: Erkennung der Buchstabengrenzen innerhalb einer Textzeile mit dem Kürzeste-Pfade-Algorithmus (oben) und vertikalen Projektionsprofilen (unten). Bei den Projektionsprofilen treten Fehler durch nicht erkannte Trenner (a) und Trennungen innerhalb von Buchstaben (b) auf.

Kosten des Pfades sind definiert als summierte Pixeldifferenzen zwischen benachbarten Pfad-pixeln. Der Pfad mit den geringsten Kosten schneidet nur selten Buchstabenpixel und eignet sich somit gut als Trenner von Buchstaben [277, 278].

Der *Kürzeste-Pfade-Algorithmus* für Graphen von *Dijkstra* [92] wird verwendet, um die Trenner zu bestimmen. Jedes Pixel entspricht einem Knoten, der mit drei Nachbarn (links, rechts und unten) verbunden ist. Die Kosten, um von einem Knoten zum nächsten zu gelangen, sind definiert als absolute Helligkeitsdifferenz dieser beiden Pixel. Der Algorithmus beginnt an einer Position in der obersten Zeile der Textregion und berechnet den Pfad bis zur untersten Zeile. Ergebnisse des Kürzeste-Pfade-Algorithmus sind in Abbildung 6.3 (oben) dargestellt. Neben den guten Ergebnissen ist ein wesentlicher Vorteil, dass keine Schwellwerte definiert werden müssen.

Der Aufwand zur Berechnung des kürzesten Pfades ist sehr hoch, falls dieser für jedes Pixel am oberen Rand der Textzeile berechnet wird. Der folgende Algorithmus reduziert den Aufwand signifikant:

1. Schätze die minimale Breite W eines Buchstabens aus der Höhe der Textregion.
2. Initialisiere jedes $\frac{W}{2}$ Pixel als mögliches *Startpixel* in der obersten Zeile der Textregion

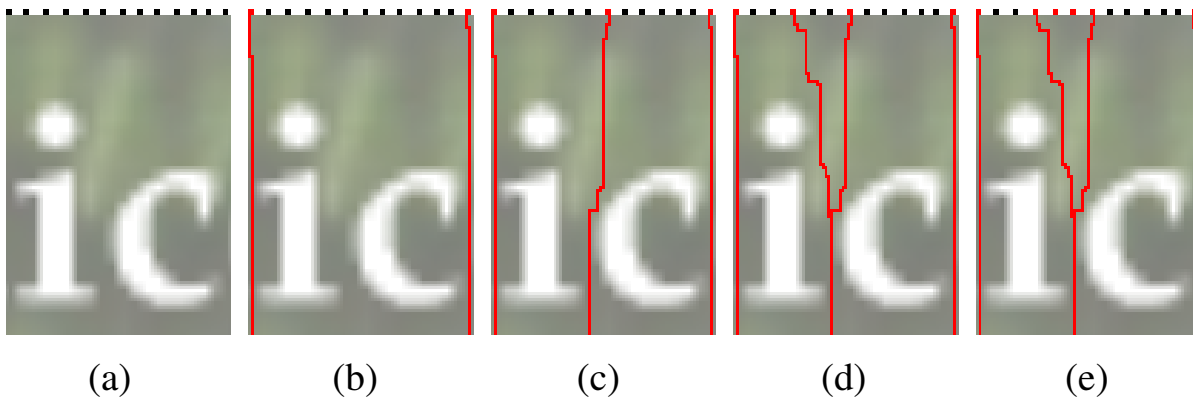


Abbildung 6.4: Optimierung des Kürzeste-Pfade-Algorithmus zur Festlegung der Trenner zwischen Buchstaben

(vgl. Abbildung 6.4 (a)).

3. Berechne den kürzesten Pfad für die Startpixel am linken und rechten Rand der Textregion (Abbildung 6.4 (b)). Alle Startpixel, deren kürzester Pfad bekannt ist, werden als *Pfadpixel* bezeichnet.
4. Wähle unter den Startpixeln das Pixel mit dem größten Abstand zu den verfügbaren Pfadpixeln (Abbildung 6.4 (c)). Der kürzeste Pfad wird berechnet und das Pixel als Pfadpixel markiert.
5. Falls ein neu berechneter kürzester Pfad mit einem anderen kürzesten Pfad zusammenfällt, ist eine weitere Berechnung des Pfades nicht mehr erforderlich, und es werden alle Startpixel zwischen den beiden Pfaden als Pfadpixel markiert. In Abbildung 6.4 (d) fallen die beiden Pfade zusammen, so dass die Startpixel zwischen den Pfaden umbenannt und nicht weiter analysiert werden müssen (e).
6. Gehe zu Schritt 4, falls weitere Startpixel verfügbar sind.

6.3.2 Identifikation der Textpixel

Die Zuordnung zu Text- oder Hintergrundpixeln erfolgt durch einen modifizierten *Region-Merging-Algorithmus*. Um diesen zu initialisieren, muss die *Textfarbe* bekannt sein. Es wird angenommen, dass es sich bei einer der beiden am häufigsten auftretenden Farben innerhalb der Textzeile um die Farbe der Buchstaben handelt. Die beiden häufigsten Farben werden durch Histogrammanalyse ermittelt, und die Textfarbe wird anhand der Position der Pixel innerhalb der Textzeile festgelegt.

In den analysierten Bildern und Videos entspricht eine der beiden am häufigsten zu beobachtenden Farben fast immer der Textfarbe. Nur in drei Prozent der analysierten Textzeilen wird die Farbe der Buchstaben nicht korrekt identifiziert. Eine fehlerhafte Textfarbe entsteht im Wesentlichen durch Kompressionsfehler, durch kleine sich bewegende Schriften oder durch zwei- bzw. mehrfarbige Buchstaben. Histogramme mit jeweils drei Bits pro Farbkanal werden zur Analyse verwendet. In den analysierten Bildern und Videos kommt die häufigste Farbe durchschnittlich in 21,9 % und die zweithäufigste Farbe in 11,3 % der Pixel vor.

Die ermittelte Textfarbe beschreibt nur einen Teil der tatsächlichen Textpixel, da durch Helligkeitsschwankungen, Rauschen und Kompressionsfehler deutliche Abweichungen der Farbe entstehen können. Eine Segmentierung ausschließlich aufgrund der Textfarbe verursacht sehr starke Segmentierungsfehler, die eine Klassifikation unmöglich machen würde.

Im zweiten Schritt werden die Blöcke zwischen zwei Trennern betrachtet und jedes Pixel als Text oder Hintergrund klassifiziert. Ein *Region-Growing-Algorithmus* bestimmt zunächst zusammenhängende Regionen einer Farbe. Anschließend werden die Regionen mit einem modifizierten *Region-Merging-Algorithmus* als Text oder Hintergrund festgelegt. Ein erweitertes Distanzmaß berücksichtigt dabei sowohl ähnliche Farben als auch die Entfernung der Regionen untereinander:

1. Jede Region kann einen der drei Zustände annehmen: *Text*, *Hintergrund* oder *undefiniert*. Alle Regionen sind zunächst undefiniert.
2. Ist die Farbe einer Region identisch mit der berechneten Textfarbe, so wird diese Region als *Text* klassifiziert.
3. Undefinierte Regionen, die an die obere oder untere Kante des Blockes angrenzen, werden als *Hintergrund* definiert.
4. Ein Distanzmaß berechnet wie folgt die Entfernungen $D_{i,j}$ zwischen jeder undefinierten Region i und allen definierten Regionen j (Text bzw. Hintergrund):

$$D_{i,j} = |C_i - C_j| + |G_i - G_j|. \quad (6.1)$$

Jede Region wird durch ihre Farbe C_i und den Schwerpunkt G_i aller Pixel der Region definiert.

5. Das Minimum von $D_{i,j}$ wird ausgewählt, und Region i wird abhängig von Region j als *Text* oder *Hintergrund* klassifiziert.
6. Der Algorithmus wird mit Schritt 4 fortgesetzt, solange weitere undefinierte Regionen verfügbar sind.

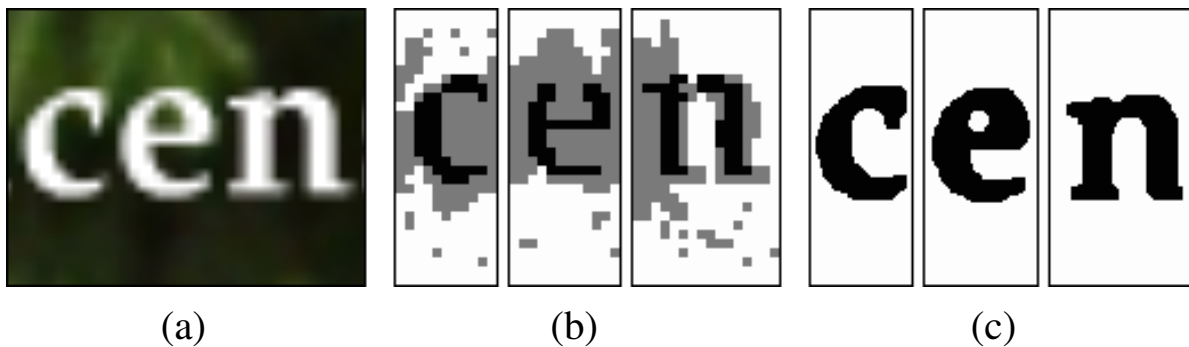


Abbildung 6.5: *Präzise Segmentierung von Textpixeln in verrauschten Bildern: Originalbild (a), initialisierte Regionen (b) und endgültige Segmentierung (c).*

Abbildung 6.5 zeigt die wesentlichen Schritte des Algorithmus am Beispiel von drei Buchstaben. Im Originalbild (a) wird deutlich, dass eine hohe Kompressionsrate zu sehr unscharfen Textpixeln führen kann. Abbildung 6.5 (b) zeigt den Zustand des Algorithmus nach Schritt 3. Die Regionen der weißen Hintergrundpixel grenzen an den oberen oder unteren Rand der Textzeile an. Die schwarzen Bereiche sind Pixel der Textfarbe und beschreiben die Buchstaben nur sehr ungenau. Die grauen Pixel sind zunächst undefiniert und werden im Verlauf des Algorithmus zu Text oder Hintergrund (Abbildung 6.5 (c)). Die Kombination von Farbinformationen und örtlichen Informationen im Distanzmaß ermöglicht eine Segmentierung in guter Qualität.

Als weiteres Verfahren zur Einteilung in Text- und Hintergrundpixel wurde der *K-Means*-Algorithmus betrachtet. Ein wesentliches Problem ist dabei die feste Anzahl von Clusterzentren. Bei zwei Zentren werden sehr viele Pixel dem jeweils falschen Zentrum zugeordnet. Um gute Segmentierungsergebnisse zu erhalten, müsste die Anzahl der Cluster von der Komplexität des Bildbereiches abhängen. Das grundsätzliche Problem, also die Entscheidung, ob ein Cluster Textpixel oder Hintergrundpixel enthält, würde durch den Algorithmus nicht gelöst. Wir haben deshalb auf eine weitere Verwendung des *K-Means*-Algorithmus verzichtet.

Im letzten Schritt wird von allen markierten Buchstabenpixeln die größte zusammenhängende Region ausgewählt. Das ist erforderlich, da vereinzelt Hintergrundpixel in Textfarbe auftreten, die sonst auch als Bestandteil eines Buchstabens gelten würden. Der Nachteil bei dieser Vorgehensweise liegt darin, dass auch Punkte auf Buchstaben entfernt werden und Umlaute nicht mehr erkannt werden können.

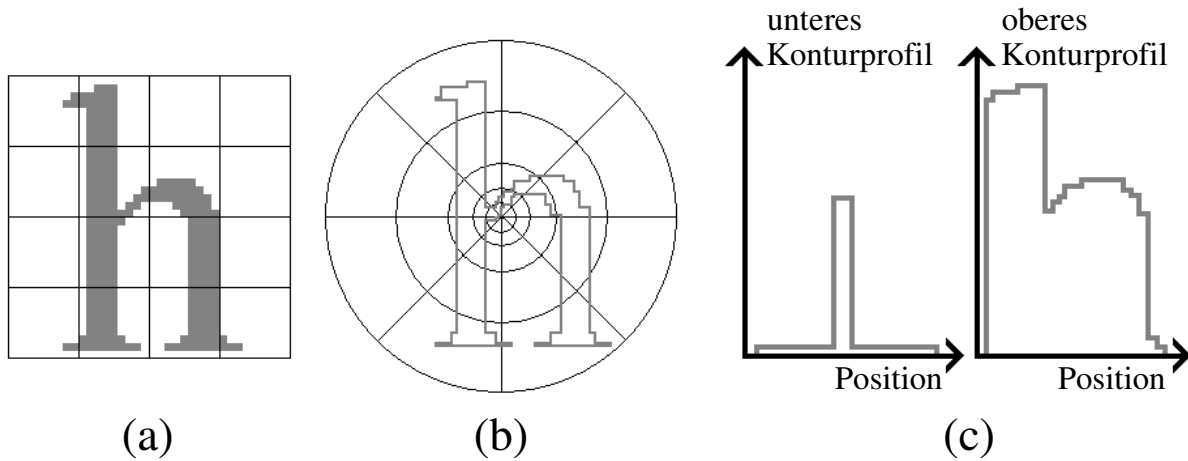


Abbildung 6.6: Merkmale zur Charakterisierung von Buchstaben beim Zoning-Algorithmus
(a), Shape-Contexts (b) und durch horizontale Konturprofile (c)

6.4 Klassifikation von Buchstaben

Vier Verfahren zur Erkennung von Buchstaben werden im Folgenden betrachtet. Dabei wird angenommen, dass eine den unbekannten Zeichen ähnliche Schriftart in der Datenbank vorhanden ist. Beim *Pattern-Matching*-Verfahren werden die Binärbilder zweier Buchstaben übereinander gelegt und der Anteil der deckungsgleichen Pixel gezählt. Ein Vorteil dieses sehr einfachen Ansatzes ist, dass im Vergleich zu konturbasierten Verfahren Segmentierungsfehler und insbesondere Unterbrechungen der Kontur weniger starke Auswirkungen auf die Klassifikation haben. Die Größe des zu analysierenden Bildes wird entsprechend der Höhe der Zeichen der Datenbank skaliert. Die Distanz $D_{Q,J}$ zweier Buchstaben ist definiert als:

$$D_{Q,J} = \frac{1}{n_x \cdot n_y} \cdot \sum_{x=1}^{n_x} \sum_{y=1}^{n_y} \begin{cases} 0 & \text{falls } Q_{x,y} = J_{x,y}, \\ 1 & \text{sonst.} \end{cases} \quad (6.2)$$

Q bezeichnet das unbekannte Zeichen, J einen Buchstaben der Datenbank. Die Distanz $D_{Q,J}$ beschreibt den Anteil der unterschiedlichen Pixel beider Buchstaben.

Beim zweiten Algorithmus handelt es sich um das sogenannte *Zoning*-Verfahren [262, 505]. Es wird ein Gitter der Größe $n \times m$ über das Zeichen gelegt (vgl. Abbildung 6.6 (a)), und die Anzahl bzw. der Anteil der Textpixel wird in jedem Gitterblock als Merkmalsvektor verwendet. Der ursprüngliche Zoning-Algorithmus des kommerziellen OCR-Systems CALERA [46] wurde verwendet, der auch Buchstaben mit deutlichen Segmentierungsfehlern erkennen soll. Star-

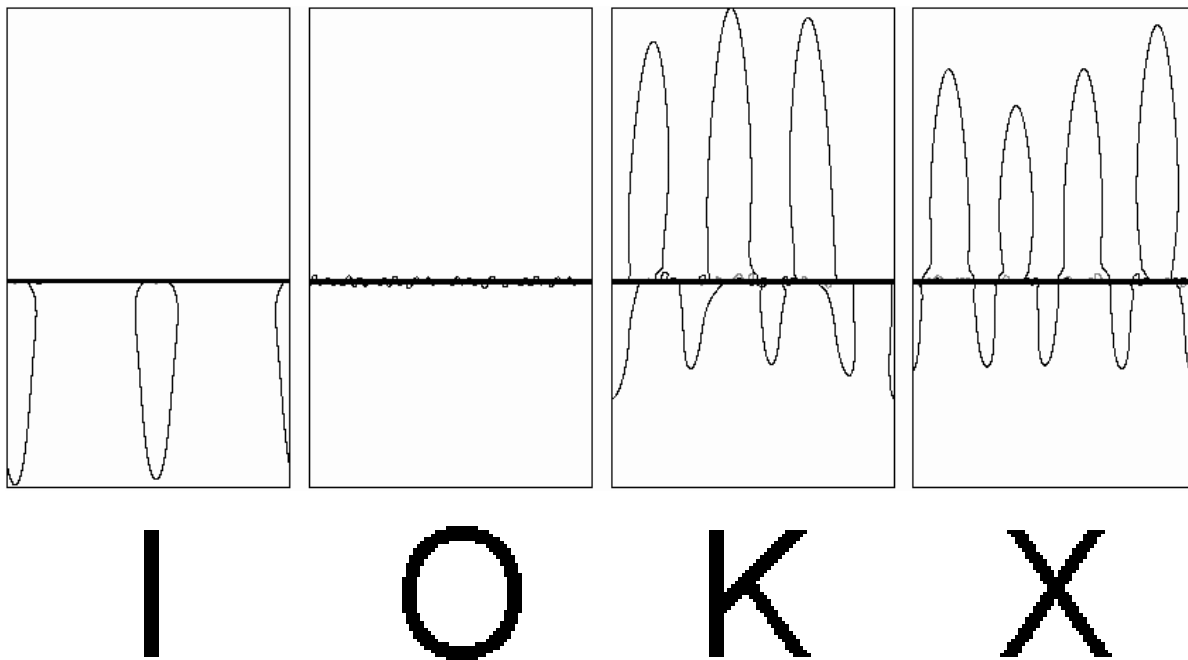


Abbildung 6.7: Beispiele für Skalenraumabbildungen von Buchstaben

ke Ähnlichkeiten mit dem Zoning-Verfahren haben sogenannte *Shape-Contexts* [30, 31, 367]. Statt einer Einteilung in rechteckige Gitterblöcke werden Kreissegmente definiert. In jedem Segment wird die Anzahl der Kantenpixel gezählt, aus denen ein charakteristischer Merkmalsvektor des Buchstabens abgeleitet wird. Abbildung 6.6 verdeutlicht die Einteilung in Regionen beim Zoning (a) und bei Shape-Contexts (b).

Ein drittes Distanzmaß nutzt *Konturprofile* zum Vergleich von Buchstaben [262, 505]. Bei einem horizontalen Konturprofil werden die oben und unten gelegenen Konturpixel eines Buchstabens analysiert, das vertikale Konturprofil berücksichtigt die Konturpixel am linken und rechten Rand. Die vier Profile definieren den Merkmalsvektor des Zeichens.

Als viertes Verfahren werden *Skalenraumabbildungen* zur Klassifikation der segmentierten Buchstaben verwendet. Viele Buchstaben haben eine sehr einfache Form mit wenigen konkaven Regionen, so dass die Verwendung der in Kapitel 5.7 vorgestellten transformierten Konturen erforderlich ist. Abbildung 6.7 verdeutlicht, dass nur durch Kombination der ursprünglichen und der transformierten Konturen ein zuverlässiger Vergleich der Skalenraumabbildungen möglich ist. Die ursprünglichen Skalenraumabbildungen der Buchstaben 'I' und 'O' sind nahezu identisch, sie unterscheiden sich jedoch deutlich bezüglich ihrer transformierten Konturen. Umgekehrt sind die transformierten Konturen der Buchstaben 'K' und 'X' sehr ähnlich.

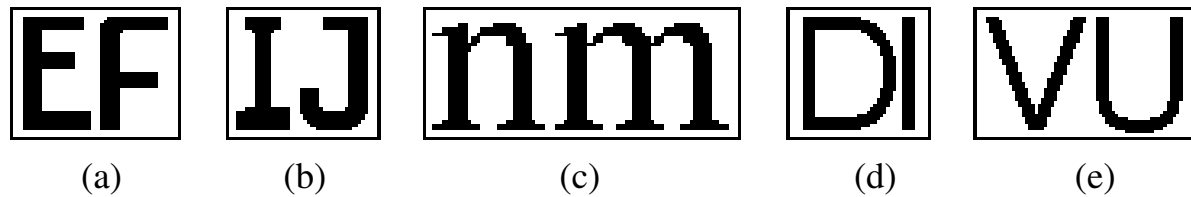


Abbildung 6.8: Beispiele für Buchstaben der unterschiedlichen Zeichensätze der Datenbank

Durch die fixe Anzahl an Abtastpunkten müssen Schriftarten mit unterschiedlichen Größen nicht gesondert betrachtet werden. Es wird angenommen, dass ein Text horizontal ausgerichtet ist. Daher sind beim Skalenraumvergleich nur Rotationen bis maximal zwanzig Grad zulässig, um leicht geneigte Buchstaben kursiver Schriftarten, auch wenn diese nicht in der Datenbank enthalten sind, erkennen zu können. Bei experimentellen Untersuchungen wurde deutlich, dass beim *Pattern-Matching* und bei *Konturprofilen* die Erkennung von nicht in der Datenbank enthaltener kursiver Zeichen häufig nicht zuverlässig möglich ist. Auch beim *Zoning*- und beim *Shape-Kontext*-Verfahren steigt bei der Erkennung kursiver Schriftarten der Anteil der fehlerhaft klassifizierten Zeichen deutlich.

6.5 Analyse der Klassifikationsergebnisse

Neben den vier vorgestellten Algorithmen (Pattern-Matching, Zoning, Konturprofile und Skalenraumabbildungen) wird die Erkennung von Buchstaben mit einer kommerziellen OCR-Software überprüft.

6.5.1 Erkennung von Buchstaben ohne Segmentierungsfehler

Geringe Änderungen der Buchstaben, verursacht durch unterschiedliche Zeichengrößen, geringe Rotationen oder Stauchungen, sollen keine großen Auswirkungen auf die Klassifikationsergebnisse haben. Besonders relevant scheint die Möglichkeit zu sein, Zeichen unterschiedlicher Schriftarten vergleichen zu können, da im Gegensatz zu eingescannten Dokumenten die Schriftarten der analysierten Bilder und Videos häufig variieren.

Die Binärbilder der Buchstaben von vier Zeichensätzen (Arial, Times, Gothic und der Zeichensatz für europäische Nummernschilder) wurden mit einer Zeichengröße von 36 als Referenz in der Datenbank gespeichert. Abbildung 6.8 zeigt einige Buchstaben der unterschiedlichen Schriftarten. Beim Nummernschild-Zeichensatz, der speziell für die automatische Erkennung

Verfahren	Anteil korrekt erkannter Buchstaben
Pattern-Matching	72,1 %
Zoning	63,2 %
Konturprofile	69,3 %
Skalenraumabbildungen	69,8 %
Skalenraumabbildungen mit transformierten Konturen	77,3 %

Tabelle 6.1: Theoretische Obergrenze der Erkennungsraten bei unterschiedlichen Zeichensätzen

entwickelt wurde, liefern alle Verfahren besonders robuste Ergebnisse. So sind die Unterschiede zwischen sonst ähnlichen Buchstaben wie 'E' und 'F' bzw. 'I' und 'J' besonders groß (vgl. Abbildung 6.8 (a) und (b)). Die Zeichen der anderen Schriftarten sind ähnlicher, so dass insbesondere beim Pattern-Matching-Verfahren deutliche Probleme auftreten (vgl. Abbildung 6.8 (c)). Konvexe Buchstaben wie z. B. 'D' und 'I' in Abbildung 6.8 (d) können mit dem einfachen Skalenraumverfahren nicht unterschieden werden. Die Ähnlichkeit einzelner Buchstaben ('V' und 'U') ist so groß, dass jedes Verfahren bei geringen Segmentierungsfehlern fehlerhafte Klassifikationsergebnisse liefert.

Zusätzlich wird überprüft, ob Zeichen in einer unbekannten Schriftart erkannt werden können. Dazu werden alle Buchstaben eines Zeichensatzes aus der Datenbank genommen und einzeln mit Hilfe der verbleibenden Zeichen der Datenbank klassifiziert. Anschließend wird die Datenbank wieder aufgefüllt und der Vergleich mit dem nächsten Zeichensatz fortgeführt. Da keine Segmentierungsfehler berücksichtigt werden, liefert das Ergebnis eine theoretische Obergrenze für die Erkennung der analysierten Buchstaben mit den jeweils drei verbleibenden Zeichensätzen. Die Tabelle 6.1 gibt die durchschnittlichen Prozentsätze an, mit denen die Buchstaben korrekt erkannt werden. Es wird deutlich, dass die Zeichensätze der Datenbank eine außerordentlich wichtige Bedeutung für die Qualität der Erkennung haben.

Beim Vergleich von Buchstaben unterschiedlicher Größe ändern sich die Erkennungsraten nur geringfügig. Erst bei einer Buchstabenhöhe von weniger als zehn Pixel steigt die Fehlerrate deutlich an. Besonders große Buchstaben beeinflussen die Ergebnisse dagegen nicht.

6.5.2 Vergleich bei fehlerhafter Segmentierung

Um die Auswirkungen von Segmentierungsfehlern zu ermitteln, werden mehrere verrauschte Varianten eines Zeichens erzeugt und analysiert. Dazu werden zufällig ausgewählte Pixel des Buchstabens durch einen lokalen Erosions- oder Dilatationsoperator mit einem Radius von



Abbildung 6.9: Beispiele stark verrauschter Buchstaben

drei modifiziert. Das ähnelt einem *Impulsrauschen* (engl. *salt and pepper noise*), das jedoch nur auf Buchstabenpixel beschränkt ist und eine 'gröbere Körnung' besitzt, so dass statt eines einzelnen Pixels jeweils ein kleiner Block verändert wird. Beispiele für besonders stark veränderte Zeichen sind in Abbildung 6.9 dargestellt.

Um die Auswirkungen des Rauschens besser vergleichen zu können, werden die Erkennungsraten entsprechend des vorherigen Abschnittes für unbekannte Zeichensätze ermittelt. Die Erkennungsraten für eine korrekte Klassifikation fallen auf 67,4 Prozent (Pattern-Matching), 62,2 Prozent (Zoning-Verfahren), 66,0 Prozent (Konturprofile), 63,9 Prozent (Skalenraumabbildungen) und 71,2 Prozent (transformierte Skalenraumabbildungen). Besonders stabil ist das Zoning-Verfahren, da es durch lokale Segmentierungsfehler nur gering beeinflusst wird. Falls die Kontur eines Zeichens wie in Abbildung 6.9 bei den Buchstaben 'D', 'w' und 'x' unterbrochen ist, können Verfahren, die die gesamte Kontur berücksichtigen, keine zuverlässigen Ergebnisse liefern.

6.5.3 Texterkennung in Bildern und Videos

Zur Erkennung der Texte in Bildern und Videos werden die automatisch segmentierten Buchstaben mit allen Buchstaben der Datenbank verglichen, und das beste Klassifikationsergebnis bestimmt den Buchstaben. Schwellwerte zum Entfernen von offensichtlich falschen Ergebnissen werden nicht verwendet. Zwanzig Bilder¹ mit komplexem Hintergrund und zehn kurze Videosequenzen² werden im Folgenden analysiert.

Das Maß für die Vollständigkeit zur Erkennung der Textzeilen liegt bei über 96 Prozent, so dass nur vereinzelt Textzeilen übersehen werden. Viele Hintergrundregionen mit starker Textur werden als Text klassifiziert, so dass die Präzision nur 63 Prozent erreicht. Durch die Überprüfung einer Region mit einfachen Heuristiken (Zeichenhöhe, Breite einer Textregion oder

¹JPEG-Kompression, Bildauflösung: 320x200 bis 800x600. Der Kompressionsfaktor wurde so gewählt, dass die Bilder auf ca. 10 Prozent im Vergleich zur unkomprimierten Dateigröße verkleinert wurden.

²MPEG-2 Video, PAL-Auflösung, 25 Bilder pro Sekunde, 6 MBit pro Sekunde.

	Kürzester-Pfade-Algorithmus	Projektionsprofile
Anteil fehlerhaft ausgewählter Farbe für Textpixel	2,9 %	2,9 %
Anteil getrennter Buchstaben	3,7 %	9,8 %
Anteil verbundener Buchstaben	2,6 %	4,7 %
Anteil fehlerhaft segmentierter Buchstaben	9,2 %	17,4 %

Tabelle 6.2: *Ergebnisse zur Segmentierung der Buchstaben*

Analyse der dominanten Textfarbe) kann ohne nennenswerte Verringerung der Zuverlässigkeit die Präzision auf 91 Prozent erhöht werden. Die untere und obere Grenze einer Textregion wird immer korrekt erkannt. Die Breite der Textregion ist häufig fehlerhaft, so dass am linken oder rechten Rand einer Textzeile durchschnittlich sechs Prozent der Buchstaben nicht erkannt werden.

Voraussetzung für die Segmentierung ist die Erkennung der korrekten Textfarbe. Bei 97,1 % aller Buchstaben wurde die Farbe korrekt identifiziert. Ein Buchstabe gilt als korrekt segmentiert, falls er nicht geteilt oder mit anderen Buchstaben verbunden ist. Zur Ermittlung der Trenner zwischen den Buchstaben werden die Ergebnisse der vertikalen Projektionsprofile mit den Ergebnissen des Kürzeste-Pfade-Algorithmus verglichen. Tabelle 6.2 fasst den Anteil der Fehler beider Verfahren zusammen. Der Kürzeste-Pfade-Algorithmus findet die Trenner zwischen den Buchstaben wesentlich zuverlässiger und reduziert die Fehlerrate von 17,4 auf 9,2 Prozent.

Die Klassifikationsergebnisse für die korrekt und fehlerhaft segmentierten Buchstaben werden unabhängig voneinander betrachtet. Mit weniger als acht Prozent korrekter Ergebnisse ist eine Erkennung der Buchstaben bei fehlerhafter Segmentierung mit keinem Klassifikationsverfahren möglich. Tabelle 6.3 gibt die Ergebnisse für die korrekt segmentierten Buchstaben in den Bildern und Videosequenzen an. Die Klassifikationsergebnisse sind für Bilder und Videos sehr ähnlich: Die Skalenraumabbildungen mit transformierten Konturen liefern die besten Klassifikationsergebnisse, dicht gefolgt von der kommerziellen OCR-Software und den Konturprofilen. Trotz der Einfachheit liefert das Pattern-Matching-Verfahren gute Ergebnisse und liegt noch vor den ursprünglichen Skalenraumabbildungen und dem Zoning.

In den Originalbildern kann das kommerzielle OCR-Softwareprodukt keine Textregionen erkennen, so dass die segmentierten Binärbilder, in denen die fehlerhaft segmentierten Buchstaben manuell entfernt wurden, für die Analyse verwendet werden. Ein objektiver Vergleich der Erkennungsraten ist nicht möglich, da das kommerzielle System als zusätzlichen Schritt einen Abgleich mit einem Wörterbuch durchführt und so einzelne nicht erkannte Buchstaben

	Bilder	Video- sequenzen
Anzahl Buchstaben	2986	1211
Pattern-Matching-Verfahren	69,1 %	77,7 %
Zoning	64,2 %	69,7 %
Konturprofile	71,2 %	82,0 %
Skalenraumabbildungen	66,9 %	78,8 %
Erweiterte Skalenraumabbildungen mit transformierten Konturen	75,6 %	88,1 %
Kommerzielles OCR-Produkt (mit Wörterbuch)	75,2 %	76,7 %
Erkennung von Textzeilen	96,6 %	97,1 %
Segmentierung mit dem Kürzeste-Pfade-Algorithmus	90,8 %	91,0 %
Gesamte Erkennungsrate mit dem erweiterten Skalenraumansatz	66,3 %	77,8 %

Tabelle 6.3: Ergebnisse zur Klassifikation der korrekt segmentierten Buchstaben

korrigieren kann. Durch die hohe Qualität der Videos ist der Anteil der Klassifikationsfehler in den Videosequenzen durchschnittlich geringer, wobei das kommerzielle System aus der höheren Qualität nur einen sehr kleinen Vorteil ziehen kann. Die in den Videos verwendeten Zeichensätze und das Wörterbuch des OCR-Systems, das für Wörter in Textdokumenten erstellt wurde, sind mögliche Ursachen für die geringeren Erkennungsraten. Abbildung 6.10 verdeutlicht die Ergebnisse der Texterkennung für ein Bild mit komplexem Hintergrund. Das Bild enthält Zeichensätze in unterschiedlicher Schriftart und Schriftgröße.

Kritisch bleibt anzumerken, dass – obwohl der erweiterte Skalenraumansatz für die analysierten Bilder und Videosequenzen bessere Ergebnisse als das kommerzielle OCR-Produkt liefert – die Fehlerraten bei allen eingesetzten Verfahren sehr hoch sind. In jedem einzelnen Schritt – also bei der Erkennung von Textzeilen, der Identifikation der Textfarbe, der Festlegung der Trenner zwischen einzelnen Buchstaben, der Segmentierung sowie der Erkennung der einzelnen Buchstaben – treten Fehler auf, die in der Summe zu den hohen Fehlerraten führen. Zudem scheinen bei den einzelnen Schritten noch deutliche Verbesserungen möglich zu sein, wie beispielsweise bei dem ausgewählten Verfahren zur Identifikation von Textzeilen. Auch die Verfahren zur Erkennung von segmentierten Buchstaben weisen zum Teil sehr hohe Fehlerraten auf. Verfahren, die beispielsweise auf der Analyse der äußeren Kontur beruhen, führen schon bei geringen Segmentierungsfehlern (Unterbrechung der Kontur) zu falsch klassifizierten Buchstaben. Zusätzliche Informationen über einzelne Buchstaben werden nicht berücksichtigt, da bei der Segmentierung Punkte auf den Buchstaben wie dem 'i', 'j' oder Umlauten entfernt werden. Auch durch den Einsatz eines Wörterbuches sind Verbesserungen zu



Abbildung 6.10: Wesentliche Schritte der Texterkennung: Originalbild (oben), automatisch erkannte Textregionen (Mitte) und segmentierter Text (unten).

erwarten, da so einzelne fehlerhafte Buchstaben korrigiert werden können.

Auffällig ist der große Qualitätsunterschied im Vergleich zur Texterkennung von eingescannten Dokumenten. Gerade durch den Einsatz kommerzieller OCR-Software für die Texterkennung in Bildern und Videos müssten signifikant bessere Ergebnisse erreicht werden können. Wesentliche Ursachen für die schlechte Qualität der kommerziellen OCR-Produkte bei Bildern und Videos sind vermutlich auf fehlende Zeichensätze und auf die Art der Aufbereitung und Segmentierung der Buchstaben zurückzuführen.

6.6 Zusammenfassung

In diesem Kapitel wurde ein Verfahren zur Segmentierung und Erkennung von Buchstaben in Bildern und Videos vorgestellt. Eine wesentliche Herausforderung sind Kompressionsartefakte und die geringe Bildauflösung. Besonders wichtig bei der Segmentierung ist die zuverlässige Erkennung der Trenner zwischen den Buchstaben, da sonst keine akzeptablen Klassifikationsergebnisse möglich sind. Zwei neue Verfahren zur Segmentierung der einzelnen Buchstaben wurden vorgestellt, die zu deutlich besseren Ergebnissen führen: Der *Kürzeste-Pfade-Ansatz* identifiziert zuverlässig *Trenner* zwischen Buchstaben, und die Erweiterung des *Region-Merging-Verfahrens*, bei dem als Distanzmaß die Entfernung zwischen Bildregionen und die Ähnlichkeit von Farben berücksichtigt werden, ermöglicht eine exakte Segmentierung. Im Durchschnitt liegen die Klassifikationsergebnisse beim Skalenraumvergleich mit transformierten Konturen deutlich über den Ergebnissen der anderen Verfahren. Im Vergleich zur Erkennung eingescannter Dokumente mittels aktueller kommerzieller OCR-Software ist eine Texterkennung in Bildern und Videos jedoch noch nicht sehr zuverlässig möglich, und es besteht weiterer Forschungsbedarf.

KAPITEL 7

Gesichtserkennung

Das menschliche Gehirn kann Gesichter in einem Bild oder Video nicht nur finden, vergleichen und identifizieren, sondern auch Emotionen und Stimmungen ablesen. Auch das Geschlecht und das ungefähre Alter lässt sich aus einem unbekannten Gesicht ableiten, obwohl die Unterschiede zwischen Gesichtern in Bezug auf die Gesichtsfarbe, Form und Anordnung der Gesichtsmarkmale wie Augen, Nase und Mund häufig sehr gering sind.

Neben der automatischen Analyse und Indexierung von Gesichtern wäre eine zuverlässige Gesichtserkennung wünschenswert, um die Interaktion zwischen Mensch und Computer zu verbessern. Für die Kommunikation unter Menschen ist es wichtig, Unsicherheit, Ablehnung oder Ironie zu erkennen. Neben dem Sprachverständnis spielen daher auch visuelle Informationen, wie beispielsweise Mimik, Gestik oder Kopfbewegungen des Gesprächspartners, eine wesentliche Rolle. Seit vielen Jahren beschränkt sich die Kommunikation mit dem Rechner im Wesentlichen auf Tastatur und Maus, zwei unnatürliche und wenig intuitiv zu bedienende Kommunikationsschnittstellen. Die Kommunikation zwischen Mensch und Maschine könnte durch eine zuverlässige und zeitnahe Lokalisierung und Analyse von Gesichtern verbessert werden, wobei für eine gute Kommunikation die Interpretation des Gesichtsausdrucks von entscheidender Bedeutung ist.

Innerhalb eines Videos liefern Gesichter besonders relevante semantische Informationen. Personen sind im Allgemeinen die Hauptakteure eines Videos, ohne die ein Verständnis der Handlung nicht möglich ist. Insbesondere für automatisch generierte Zusammenfassungen und für die computergestützte Inhaltsadaption von Videos liefern Gesichter einen wichtigen Hinweis auf relevante Kameraeinstellungen und Bildbereiche.

In diesem Kapitel werden Algorithmen zur automatischen Lokalisierung und Erkennung von Gesichtern vorgestellt. Dabei sollen die Verfahren Informationen liefern, um zusätzliche semantische Informationen aus Videos zu gewinnen. Obwohl ein umfangreicher Überblick über existierende Vorarbeiten zur Gesichtserkennung gegeben wird, soll das Forschungsgebiet der Gesichtserkennung nicht in voller Tiefe behandelt werden. Das würde den Umfang der Arbeit sprengen und einen anderen Schwerpunkt in dieser Arbeit setzen. Für die Lokalisierung und Erkennung von Gesichtern werden zwei bekannte Verfahren ausgewählt und kurz vorgestellt. Diese liefern ausreichend genaue Gesichtsinformationen für die weitere semantische Analyse von Videos. Zusätzlich werden die Gesichtsinformationen in weiteren Anwendungen wie beispielsweise der Adaption von Videos oder der automatischen Erzeugung von Zusammenfassungen in den Kapiteln 8 und 9 genutzt.

Im folgenden Abschnitt werden zunächst die besonderen Herausforderungen erläutert, die an eine Gesichtserkennung gestellt werden. Eine Klassifikation der Verfahren zur Gesichtserkennung, die wir im Folgenden in modellbasierte und konnektionistische Verfahren untergliedern, erfolgt in Abschnitt 7.2. In Abschnitt 7.3 wird die Gesichtserkennung für Videos als ein dreistufiger Prozess vorgestellt, der sich aus der Lokalisierung einer Gesichtsregion, der Segmentierung (Feinlokalisierung) und Normalisierung des Gesichtes sowie der eigentlichen Gesichtserkennung zusammensetzt. Für den ersten und dritten Schritt wird auf bekannte Verfahren zurückgegriffen. Der zweite Schritt – die Feinlokalisierung und Aufbereitung des Gesichtes – ist für eine zuverlässige Erkennung besonders wichtig. Ein neuer Algorithmus zur genauen Segmentierung und Normalisierung des Gesichtes wird vorgeschlagen, bei dem Skalierungsunterschiede, Rotationen, der Kontrast und Beleuchtungsunterschiede ausgeglichen werden. Innerhalb der experimentellen Ergebnisse in Abschnitt 7.4 werden neue Möglichkeiten aufgezeigt, um weitere semantische Informationen aus den erkannten Gesichtern abzuleiten. Beispielsweise werden besonders relevante Personen erkannt, die Anzahl der Personen im Video ermittelt, Personengruppen in Videos identifiziert oder Bildbereiche erkannt, in denen sich Personen üblicherweise aufhalten.

7.1 Anforderungen an Algorithmen zur Gesichtserkennung

Eine besondere Herausforderung für Algorithmen zur Gesichtserkennung liegt in der großen Anzahl von Faktoren, die das Aussehen eines Gesichtes beeinflussen. Ein wesentlicher Faktor ist die *Richtung der Beleuchtung*, die zu Schatten und starken Helligkeits- oder Texturveränderungen in einzelnen Gesichtsregionen führen kann [29, 178, 179]. Die *Art der Beleuchtung*

(Tageslicht, Kunstlicht oder farbiges Licht) hat starken Einfluss auf die Gesichtsfarbe. Die *Drehung oder Neigung des Kopfes* und die *Mimik einer Person* betonen oder verdecken einzelne Gesichtsmarkmale. *Skalierungsunterschiede* erfordern eine exakte Lokalisierung und Anpassung der Gesichtsgröße vor dem eigentlichen Vergleich. *Äußere Veränderungen*, die durch Schminke, Kleidungsstücke (Mütze, Schal oder Brille) und durch eine Änderung der Frisur oder des Bartes hervorgerufen werden, können einen ganz anderen Eindruck eines Gesichtes erzeugen. Auch *Verdeckungen* durch andere Objekte beeinflussen die Möglichkeit für eine korrekte Erkennung. Durch *natürliches Altern* ändert sich ein Gesicht im Laufe der Jahre, was beispielsweise in Reisepässen bei der computergestützten Verifikation eines Gesichtes berücksichtigt werden muss.

Jeder einzelne Einflussfaktor kann zu deutlichen Unterschieden zwischen zwei Bildern einer Person führen. Andererseits ist es möglich, dass bei Geschwistern und insbesondere bei Zwillingen zwei Gesichter so ähnlich sind, dass auch Menschen diese nur mit Mühe unterscheiden können. Ein Algorithmus zur Gesichtserkennung muss die individuellen Gesichtsmarkmale stärker als die Unterschiede zwischen den äußeren Einflussfaktoren wie Beleuchtung, Kleidung oder einer Drehung des Kopfes berücksichtigen.

7.2 Verfahren zur Gesichtserkennung

Unterschiedliche Ziele können bei der Analyse von Gesichtern in Bildern und Videos verfolgt werden. Bei der *Lokalisierung einer Gesichtsregion* (engl. *face detection*) wird die Position eines oder mehrerer Gesichter im Bild bestimmt. Innerhalb der Gesichtsregion können *spezielle Gesichtsmarkmale* wie Augen, Nase oder Mund bestimmt werden [101, 175, 595].

Die *Gesichtserkennung* (engl. *face recognition*) geht noch einen Schritt weiter und identifiziert eine Person in einem Bild durch Vergleich mit Bildern einer Datenbank [16]. Bei der *Authentifizierung von Gesichtern* wird überprüft, ob ein Gesicht eine bekannte Person zeigt [249, 291, 499, 506]. Semantische Informationen über ein Gesicht liefert die Analyse des *Gesichtsausdrucks* [95, 118, 129, 180, 312].

Für die computergestützte Inhaltsanalyse von Videos sind insbesondere Informationen über die Gesichtsregionen und die Gesichtserkennung relevant. Echtzeitanforderungen, die für die Videoüberwachung (engl. *video surveillance*) erforderlich sind, spielen bei der Analyse von Filmen in Videoarchiven eine untergeordnete Rolle [172, 477, 597].

Wegen der großen Bedeutung von Gesichtern wurden in den letzten Jahren viele Verfahren zum Auffinden von Gesichtsregionen und zur Erkennung eines Gesichtes entwickelt [24, 75,

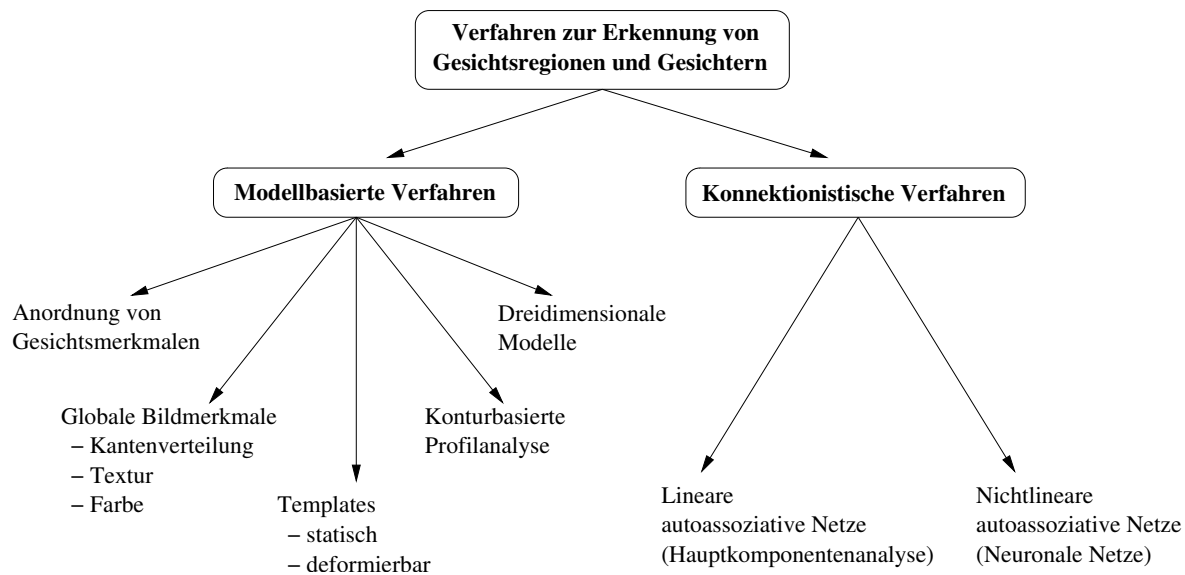


Abbildung 7.1: Klassifikation von Algorithmen zum Auffinden und Erkennen von Gesichtern

149, 150, 585]. Mehrere umfangreiche Publikationen vergleichen und beurteilen die unterschiedlichen Verfahren, die als modellbasierte oder konnektionistische Verfahren kategorisiert werden können [203, 333, 447, 541, 566, 587].

Bei den *modellbasierten Verfahren* werden Gesichter durch Regeln und Merkmale definiert, und ein Algorithmus überprüft, ob ein gegebenes Bildmuster diesen Regeln entspricht [192, 193, 254, 446]. Markante Gesichtsmerkmale, wie beispielsweise Augen, Nase und Mund, werden identifiziert, und die Beziehungen zwischen diesen Merkmalen definieren ein Gesicht. Die Klassifikationsergebnisse hängen von den ausgewählten Merkmalen und der Genauigkeit ab, mit der diese identifiziert werden können. Ein Vorteil der modellbasierten Verfahren besteht darin, dass sie auch bei Größenänderungen und Beleuchtungsunterschieden einsetzbar sind.

Die *konnektionistischen Verfahren*, zu denen beispielsweise neuronale Netze oder die Hauptkomponentenanalyse zählen, leiten die charakteristischen Merkmale eines Gesichtes selbstständig aus einer Trainingsmenge mit Gesichtsbildern ab. Abbildung 7.1 gibt einen Überblick über die im Folgenden vorgestellten Verfahren zur Lokalisierung von Gesichtsregionen und zur Erkennung von Gesichtern.

7.2.1 Modellbasierte Verfahren

Bei den *modellbasierten Verfahren* werden die Merkmale eines Gesichtes durch Regeln beschrieben [59, 144]. Da der Abstand von Augen, Nase und Mund für jede Person genau

messbar ist, kann die *Anordnung der Gesichtsmerkmale* zur Erkennung eines Gesichtes eingesetzt werden [60, 62, 290, 562]. Durch eine Drehung des Kopfes oder eine Änderung der Mimik verschiebt sich das Verhältnis der Merkmale im Bild, so dass die wesentliche Herausforderung dieses Ansatzes in der geeigneten Auswahl an Regeln liegt.

Auch *allgemeine globale Merkmale* des Gesichtes, wie beispielsweise die Kantenverteilung, Textur oder Farbe, eignen sich zur Beschreibung eines Gesichtes. Starke *Kanten* treten häufig in Bereichen der Augen, der Augenbrauen oder des Mundes auf. Durch Gruppierung zu Kantenregionen und der Aggregation benachbarter Regionen können Bildbereiche mit Gesichtern erkannt werden [81, 151]. Die Orientierung der Kanten innerhalb der einzelnen Gesichtsregionen gibt weitere Hinweise auf ein Gesicht [574, 596]. Leung et al. verwenden zur Lokalisierung von Gesichtsregionen fünf Gesichtsmerkmale (zwei Augen, zwei Nasenflügel und den Übergang von Nase und Mund) und prüfen, ob die Anordnung und Form der ermittelten Merkmale im Bild einem Gesicht entspricht [304].

Texturen ermöglichen die Identifikation von Gesichtsregionen, wobei im Wesentlichen zwischen Haut, Haaren und sonstigen Regionen unterschieden wird [18, 132, 336]. Die Textur der einzelnen Bildbereiche wird ermittelt und bei entsprechender Anordnung dieser Regionen als Gesichtsregion definiert. Eine Analyse von Texturen hat den Vorteil, dass auch gedrehte und skalierte Gesichter erkannt werden können und dass eine Verdeckung einzelner Gesichtsbereiche nur geringe Auswirkungen auf die Klassifikationsergebnisse hat.

Ein weiteres allgemeines globales Merkmal ist die *Gesichtsfarbe*, die in vielen Verfahren zur Auswahl möglicher Gesichtsregionen eingesetzt wird [176, 398, 564]. Trotz einer Vielzahl von Hauttönen unterscheidet sich die Gesichtsfarbe im Wesentlichen nur durch ihre Helligkeit und nicht durch ihre Chrominanzwerte. Damit ist es möglich, Farbintervalle für Gesichtspixel zu definieren und eine effiziente Pixelauswahl für Gesichtsregionen zu treffen. Unter Verwendung einer umfangreichen Bildsammlung haben Jones et. al die Farben von fast einer Milliarde Gesichtspixeln analysiert [240]. Für die untersuchten Bilddaten liefern Histogramme zur Beschreibung der Farben einer Gesichtsregion besonders zuverlässige Klassifikationsergebnisse. Probleme treten bei unnatürlichen Beleuchtungsverhältnissen auf, die beispielsweise durch bunte Lampen oder einen Sonnenuntergang entstehen. Die meisten Ansätze verwenden die Gesichtsfarbe zur Auswahl möglicher Gesichtsregionen und überprüfen diese anschließend mit einem weiteren Verfahren [236, 443, 466, 565].

Bei einem Vergleich mit *statischen Templates* werden Gesichtsmuster in einem Bild gesucht, indem die Korrelation zwischen dem unbekannten Bildausschnitt und dem Gesichtsmuster berechnet wird [100, 174, 353, 442, 445]. *Deformierbare Templates*, in denen die Anordnung

der einzelnen Gesichtsmerkmale durch elastische Modelle abgebildet wird, ermöglichen auch die Lokalisierung von Gesichtern mit unterschiedlicher Mimik [293, 420, 578]. Die Ähnlichkeit zweier Gesichter hängt von den Übereinstimmungen der einzelnen Merkmale und den Beziehungen zwischen den Merkmalen ab. Ein Gesicht kann durch einen elastischen Graphen abgebildet werden, in dem die charakteristischen Gesichtsmerkmale den Knoten im Graphen entsprechen [61, 274, 349, 548]. Beim Vergleich wird ein Graph so lange verändert, bis beide Graphen möglichst ähnlich sind. Die Ähnlichkeit zweier Gesichter wird durch den Umfang der Änderung der inneren Struktur der Graphen definiert. Auch äußere Veränderungen wie Brille oder Bart können bei elastischen Graphen berücksichtigt werden [548]. *Morphing* ist eng mit deformierbaren Templates verwandt, da Parameter gesucht werden, um ein Gesicht in ein anderes zu überführen [35, 358]. Ein wesentlicher Vorteil der deformierbaren Templates ist die Robustheit bei Beleuchtungsänderungen und bei einer Änderung des Gesichtsausdrucks.

Die Genauigkeit der Algorithmen, die das *Profil* eines Gesichtes analysieren, ist deutlich geringer als die der vorher beschriebenen Verfahren [577]. Markante Punkte auf dem Profil eines unbekannten Gesichtes definieren einen Merkmalsvektor, der mit anderen Profilen verglichen wird. Durch Kombination von frontalen Aufnahmen und Aufnahmen im Profil können *dreidimensionale Modelle* eines Gesichtes ermittelt werden [39, 50, 54]. Das Profil liefert die Tiefeninformation des Gesichtes, die frontale Aufnahme die Textur und genaue Position von Augen und Mund [171, 308]. Die Kombination von Tiefenkarte und Textur ermöglicht es, den Einfluss der Beleuchtung und beliebige Rotationen eines Gesichtes auszugleichen [34, 334, 402]. Modellbasierte Verfahren können gut miteinander kombiniert werden, indem jedes einzelne Verfahren als Filter interpretiert wird, der Bildbereiche entfernt, in denen mit Sicherheit kein Gesicht enthalten ist [182, 410, 546]. Durch die iterative Anwendung mehrerer Filter sind Algorithmen zur Gesichtserkennung in Echtzeit möglich [593, 594].

7.2.2 Konnektionistische Verfahren

Die *konnektionistischen Verfahren* analysieren Bilder einer Trainingsmenge, erkennen automatisch die relevanten Merkmale dieser Trainingsmenge und verwenden sie zur Analyse eines unbekannten Bildes. Für die Gesichtserkennung nehmen *konnektionistische Modelle* (engl. *connectionist model*), die im Rahmen der Psychologie zur Abbildung des menschlichen Lernens entwickelt wurden, eine zentrale Rolle ein [140, 141, 523]. Mentale Vorgänge oder Verhaltensphänomene werden ähnlich den Neuronen im menschlichen Gehirn mit vernetzten Knoten modelliert. Jeder Knoten erhält als Eingabe Daten von anderen Knoten, fasst diese

zusammen und erzeugt eine Ausgabe, die weiteren Knoten als Eingabe dient. Das Lernen oder Trainieren eines Netzes erfolgt durch eine Veränderung der Gewichte der Verbindungen zwischen den einzelnen Knoten.

Die zur Beschreibung eines Gesichtes besonders relevanten Informationen werden aus den Bilddaten der Trainingsmenge automatisch abgeleitet und stimmen nicht mit den Gesichtsmerkmalen der modellbasierten Verfahren überein. Unter den konnektionistischen Modellen werden *lineare autoassoziative Netze* (engl. *linear autoassociative network*) [9, 270, 400, 461, 521] und *nichtlineare autoassoziative Netze* eingesetzt [94, 95, 143, 332]. Zu den bekanntesten konnektionistischen Verfahren im Rahmen der Gesichtserkennung zählen *neuronale Netze* (engl. *neural net*) und die *Hauptkomponentenanalyse* (engl. *principal component analysis*), die auch im Folgenden zur Lokalisierung von Gesichtsregionen bzw. zur Klassifikation von Gesichtern eingesetzt werden.

Lineare autoassoziative Netze

Bei linearen autoassoziativen Netzen handelt es sich um ein statistisches Verfahren, das relevante Merkmale aus einer Trainingsmenge automatisch bestimmt, diese mit einer linearen Funktion transformiert und die erlernten Merkmale zur Erkennung von Gesichtsregionen oder Gesichtern verwendet [401, 513, 515, 522]. Die Idee basiert auf der *Hauptkomponentenanalyse* [152, 239], die auch unter dem Namen *Karhunen-Loève-Transformation* oder *Hotelling-Transformation* bekannt ist [252, 269, 330]. Als Eingabe dienen die Helligkeitswerte der Pixel der Gesichtsregion, die als Merkmale einen Punkt in einem vieldimensionalen Raum definieren.

Um aus den Gesichtsdaten einer Trainingsmenge die wesentlichen Faktoren zu extrahieren, wird eine Hauptachsentransformation durchgeführt. Als Faktoren werden die Eigenvektoren der Kovarianzmatrix verwendet, die einen Unterraum, den sogenannten *Gesichtsraum* (engl. *face space*) aufspannen, der alle Gesichtsbilder der Trainingsmenge enthält. Durch Linearkombination der Eigenvektoren ist es möglich, alle Gesichter der Trainingsmenge verlustfrei darzustellen. Wird nur eine Teilmenge der Eigenvektoren zur Rekonstruktion eines Bildes ausgewählt, so wird bei der verlustbehafteten Annäherung der mittlere quadratische Fehler minimiert. Geeignete Eigenvektoren für die Rekonstruktion zeichnen sich durch hohe Eigenwerte aus. Die Eigenvektoren werden auch *Eigenbilder* (engl. *eigenpicture*) oder *Eigengesichter* (engl. *eigenface*) genannt und definieren die Merkmale, aus denen ein Gesicht erzeugt wird. Eigenvektoren beschreiben keine einzelnen Gesichtsm Merkmale wie Augen, Nase oder Mund, sondern kombinieren Informationen aus allen Bereichen des Gesichtes.

Um zu überprüfen, ob es sich bei einer unbekannten Bildregion um ein Gesicht handelt, wird

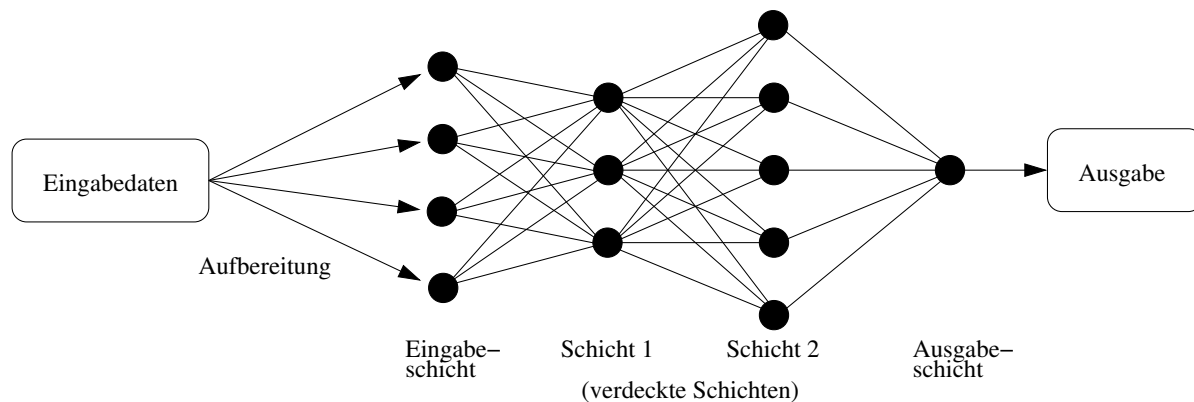


Abbildung 7.2: Beispiel für die Struktur eines neuronalen Netzes

diese Region in den durch die Eigenvektoren aufgespannten Unterraum projiziert. Dazu wird die Matrix, die den Gesichtsraum aufspannt, mit den als Vektor gespeicherten Gesichtsdaten multipliziert. Die Projektion ändert Gesichtsregionen nur geringfügig, Regionen ohne Gesichter jedoch signifikant. Ein Vergleich der ursprünglichen mit der transformierten Bildregion liefert ein Maß zur Lokalisierung von Gesichtsregionen [269, 514].

Sirovich und Kirby haben eines der ersten Verfahren zur Analyse von Gesichtern mit Hilfe der Hauptkomponentenanalyse vorgeschlagen, wobei der Schwerpunkt ihres Verfahrens in der Entwicklung eines effizienten Verfahrens zur Kodierung von Gesichtern liegt [269, 461]. Ein Gesicht wird durch Gewichte definiert, welche die Eigenvektoren der Bilder der Trainingsmenge kombinieren. Für die Kompression der Gesichtsdaten werden nur Eigenvektoren mit hohen Eigenwerten berücksichtigt, da diese die allgemeine Form eines Gesichtes beschreiben. Um beispielsweise das Geschlecht einer Person zu bestimmen, reicht eine Analyse der ersten beiden Eigenvektoren häufig aus [401]. Eigenvektoren mit niedrigeren Eigenwerten liefern detailliertere Informationen, die insbesondere für die Gesichtserkennung erforderlich sind.

Nichtlineare autoassoziative Netze

Im Fall von *nichtlinearen autoassoziativen Netzen* sind die Eingabeschichten (engl. *input layer*) nicht direkt mit den Ausgabeschichten (engl. *output layer*) verbunden, sondern erhalten ihre Daten über verdeckte Schichten (engl. *hidden layer*). Ursprünglich wurde diese Netzart von Webos entwickelt [542] und Jahre später von mehreren anderen Autoren fast zeitgleich wieder aufgegriffen [103, 411, 441]. Unter den nichtlinearen Netzen werden neuronale Netze und Support-Vector-Maschinen eingesetzt [181, 201, 219, 241, 251]. Abbildung 7.2 zeigt beispielhaft eine mögliche Struktur eines neuronalen Netzes.

Das Trainieren eines Netzes erfolgt in zwei Schritten. Bei der Initialisierung werden für alle Knoten einer Schicht die Eingabedaten mit einer nichtlinearen Funktion transformiert und summiert. Als Eingabedaten können die Pixel des Bildbereiches oder abgeleitete Merkmalsvektoren verwendet werden [128]. Die Ausgabe eines Knotens dient als Eingabe für die nächste Schicht. Im zweiten Schritt werden die Bilder der Trainingsmenge mit dem neuronalen Netz analysiert und die Klassifikationsfehler ermittelt. Um die Fehler zu korrigieren, wird das Netz in umgekehrter Richtung Schicht für Schicht durchlaufen und die fehlerhaften Daten durch das Netz geleitet. Die Gewichte der Matrix werden angepasst, so dass der durchschnittliche quadratische Fehler minimiert wird. Je stärker ein Knoten für einen Fehler verantwortlich und je höher dieser Fehler ist, desto stärker wird das Gewicht eines Knotens verändert.

Man spricht von einem *komprimierenden Netzwerk* (engl. *compression network*), falls das Netz weniger verdeckte Knoten als Eingangsknoten enthält. Durch die verdeckten Knoten werden die Daten kompakt in einem kleineren Unterraum abgebildet, wobei die relevanten Gesichtsinformationen erhalten bleiben. Redundante bzw. korrelierte Daten werden statistisch erfasst und ausgefiltert. Wird die Transformation mit einer linearen Funktion durchgeführt, entspricht das Verfahren der Hauptkomponentenanalyse, und die verdeckten Knoten beschreiben den gleichen Unterraum wie die Eigenvektoren mit den höchsten Eigenwerten [523].

Mehrere Systeme zur Erkennung von Gesichtern mit nichtlinearen autoassoziativen Netzen wurden erfolgreich entwickelt [110, 165]. Cottrell et al. verwenden ein dreischichtiges Netz mit 16 verdeckten Knoten und jeweils 64 Ein- und Ausgabeknoten zur Erkennung von Gesichtsregionen [96]. Obwohl ein nichtlineares Netz verwendet wird, spannen die ersten dreizehn verdeckten Knoten denselben Unterraum wie die Eigenvektoren der Hauptkomponentenanalyse auf. Im Vergleich zu den Eigenvektoren ist die Varianz innerhalb der Knoten jedoch gleichmäßiger verteilt [96, 97]. In weiteren Experimenten verwenden die Autoren 80 verdeckte und 4096 Ein- und Ausgabeknoten [94, 143]. Für eine Lernmenge mit 64 Gesichtern liegt die Fehlerrate für die Gesichtserkennung bei drei Prozent. Wird ein Fünftel des Gesichtes verdeckt, so steigt der Fehler um 3 bis 29 Prozent, wobei der Bereich der Augen besonders relevant ist und die Kinnregion die geringste Bedeutung hat. Helligkeitsänderungen der Gesichtsbilder erhöhen die Fehlerrate um 7 Prozent.

In den ersten Ansätzen mit neuronalen Netzen konnten nur Gesichter einer festen Größe erkannt werden [4, 245, 424]. Soulie et al. haben mehrere Netze trainiert, von denen jedes einzelne Gesichter einer festen Größe erkennt [472]. Alternativ ist eine Skalierung des Bildes und die Analyse des Bildes in allen Skalierungsstufen möglich [435].

Der Unterschied zwischen linearen und nichtlinearen autoassoziativen Netzen ist in Bezug auf

die Ergebnisse und die interne Repräsentation der Daten sehr gering [523]. Die Initialisierung der nichtlinearen autoassoziativen Netze ist mit deutlich höherem Aufwand verbunden, wobei das Ergebnis eine Annäherung der Hauptkomponentenanalyse ist. Für binäre Entscheidungsprobleme wie der Frage, ob eine Bildregion ein Gesicht zeigt oder ob ein gefundenes Gesicht weiblich oder männlich ist, sind nichtlineare Netze gut geeignet, da das Netz nur einmal erzeugt werden muss.

7.3 Lokalisierung und Erkennung von Gesichtern in Videos

Die Lokalisierung und Erkennung soll im Folgenden auf *frontale Gesichter* beschränkt werden. Diese haben in Videos eine besonders starke semantische Bedeutung: Bei einer Suche in Videoarchiven werden häufig Kameraeinstellungen, in denen Personen frontal abgebildet sind, bevorzugt betrachtet. Auch innerhalb von Zusammenfassungen von Videos oder in adaptierten Videos sind frontale Gesichtsaufnahmen im Allgemeinen besonders aussagekräftig.

Die Lokalisierung und Erkennung frontaler Gesichter erfolgt in drei Schritten. Zunächst analysiert ein neuronales Netz die Bilder eines Videos und ermittelt alle Regionen, in denen jeweils ein frontales Gesicht abgebildet ist. Die Größe und Position einer Gesichtsregion ist für die Gesichtserkennung zu ungenau, so dass in einem zweiten Schritt eine exakte Segmentierung (Feinlokalisierung) des Gesichtes erfolgt. Unter Verwendung modellbasierter Verfahren wird die Position der Augen bestimmt, so dass das Gesicht passend gedreht und auf eine einheitliche Größe skaliert werden kann. Die eigentliche Gesichtserkennung erfolgt durch ein lineares autoassoziatives Netz.

7.3.1 Lokalisierung von Gesichtsregionen

Eine der zentralen Arbeiten im Bereich der Lokalisierung von Gesichtsregionen geht auf Rowley et al. zurück, die ein dreischichtiges neuronales Netz einsetzen [435, 436, 437, 438]. Um einheitlich skalierte Gesichtsregionen zu erhalten, werden in den Bildern der Trainingsmenge die Positionen der Augen, der Nase und des Mundes markiert. Eine 20×20 Pixel große Gesichtsregion, die Pixel für Pixel über das zu analysierende Bild geschoben wird, definiert den Eingabevektor. Die Ausgabeschicht aggregiert alle Daten der unbekannten Bildregion zu einem Wert.

Um größere Gesichter zu lokalisieren, wird das Bild schrittweise verkleinert, und jedes skalierte Bild wird erneut mit dem neuronalen Netz analysiert. In Bereichen mit Gesichtern findet



Abbildung 7.3: Beispiele für die Erkennung von Gesichtsregionen

das neuronale Netz auf den unterschiedlichen Skalierungsstufen mehrere überlappende Gesichtsregionen, die zu einer einzigen Gesichtsregion aggregiert werden. Um leicht geneigte Gesichter zu finden, schlagen Rowley et al. vor, den unbekannten Bildbereich zunächst passend auszurichten. In hochauflösenden Bildern liegen die Erkennungsraten abhängig von der Qualität des Bildmaterials zwischen 85 und 95 Prozent und sinken durch den Ausgleich der Rotation auf unter 79 Prozent [436]. Um ein Absinken der Erkennungsrate zu verhindern, wird für die Analyse der Videos das gesamte Bild um fünfzehn Grad nach links und nach rechts gedreht, so dass auch leicht geneigte Gesichter gefunden werden. Abbildung 7.3 zeigt Beispiele für automatisch gefundene Gesichter innerhalb einer Dokumentation.

Der Aufwand zur Initialisierung eines neuronalen Netzes ist sehr hoch, da eine umfangreiche Lernmenge manuell zusammengestellt und aufbereitet werden muss [313, 373]. Obwohl der größte Teil der Gesichtsregionen gefunden wird und nur vereinzelt Fehlklassifikationen auftreten [373], ist die Position und Größe der Gesichtsregionen relativ ungenau, so dass ein Erkennungsalgorithmus für diese Gesichtsregionen sehr schlechte Ergebnisse liefert.

7.3.2 Segmentierung eines Gesichtes

Um gute Klassifikationsergebnisse für die Gesichtserkennung zu ermöglichen, ist eine exakte Segmentierung des Gesichtes notwendig. Im Folgenden wird ein Verfahren zur Feinlokalisierung und Normalisierung des Gesichtes vorgeschlagen, das Unterschiede in Bezug auf Rotation, Skalierung, Beleuchtung und Kontrast ausgleicht.

Durch die Aggregation der lokalisierten Gesichtsregionen über mehrere unterschiedlich skalierte Bilder weicht die gefundene Gesichtsregion zum Teil deutlich von dem tatsächlichen Gesicht ab. Um sicherzustellen, dass ein Gesicht vollständig in der Gesichtsregion enthalten

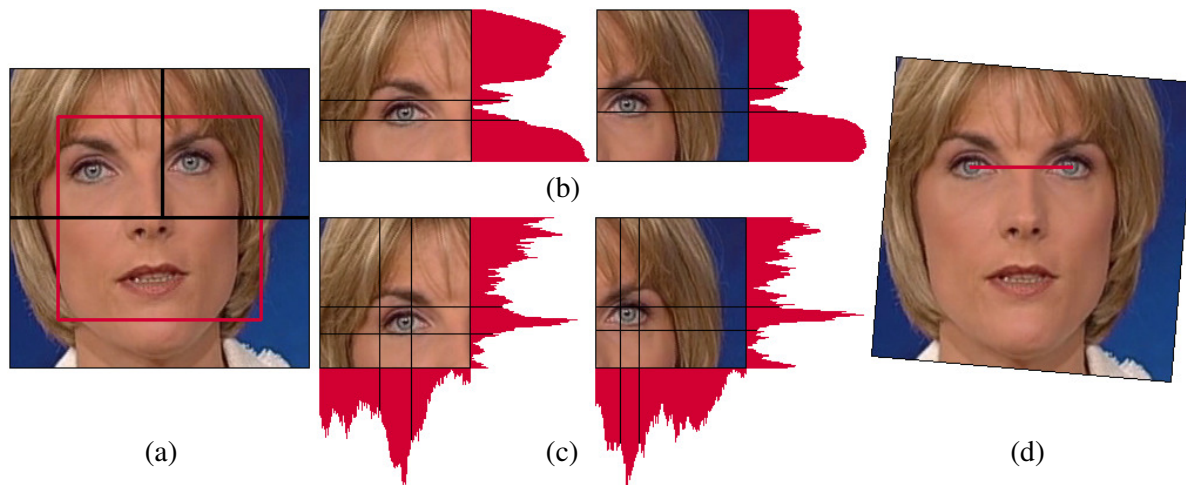


Abbildung 7.4: Ermittlung der Augen innerhalb einer Gesichtsregion: 50% vergrößerte Gesichtsregion (a), horizontale Konturprofile mit Helligkeitswerten (b), Konturprofile mit Differenzwerten benachbarter Pixel (c), Ausgleich der Rotation (d).

ist, wird der erkannte Gesichtsbereich um fünfzig Prozent vergrößert. Bis zu 15 Grad seitlich geneigte Gesichter werden durch das neuronale Netz gefunden, so dass das Gesicht zunächst passend ausgerichtet werden muss. Zur Ermittlung der Neigung des Kopfes eignen sich modellbasierte Verfahren, da diese die Positionen der einzelnen Gesichtsmarkmale genau bestimmen. Insbesondere die Augen liefern wichtige Informationen zum Ausgleich der Rotation.

Abbildung 7.4 verdeutlicht, dass Projektionsprofile des linken und rechten oberen Quadranten eine genaue Lokalisierung der Augen ermöglichen. In einem horizontalen auf Helligkeitswerten basierenden Profil wird die Augenregion (dunkler Bereich) durch das lokale Minimum definiert (Abbildung 7.4 b). Profile, die Differenzwerte benachbarter Pixel summieren (hoher Kontrast), ermöglichen sowohl in vertikaler als auch in horizontaler Richtung die Lokalisierung der Augen (Abbildung 7.4 c). Innerhalb des durch die Profile spezifizierten Bereiches wird der Mittelpunkt der Augen durch ein Pattern-Matching-Verfahren bestimmt, indem ein passend skaliertes Prototyp eines Auges, der aus zehn hochauflösenden Beispielbildern gewonnen wurde, über diesen Bereich geschoben und die minimale Differenz ermittelt wird. Abbildung 7.4 verdeutlicht die Vorgehensweise, indem zuerst die Gesichtsregion um 50 Prozent vergrößert wird, der Bereich der Augen durch Profile eingegrenzt und anschließend das Zentrum des Auges durch das Pattern-Matching-Verfahren spezifiziert wird. Das Gesicht wird gedreht, so dass beide Augen auf einer waagrechten Linie liegen.

Skalierungsunterschiede zweier Gesichter werden anhand des Augenabstandes normiert. Ein

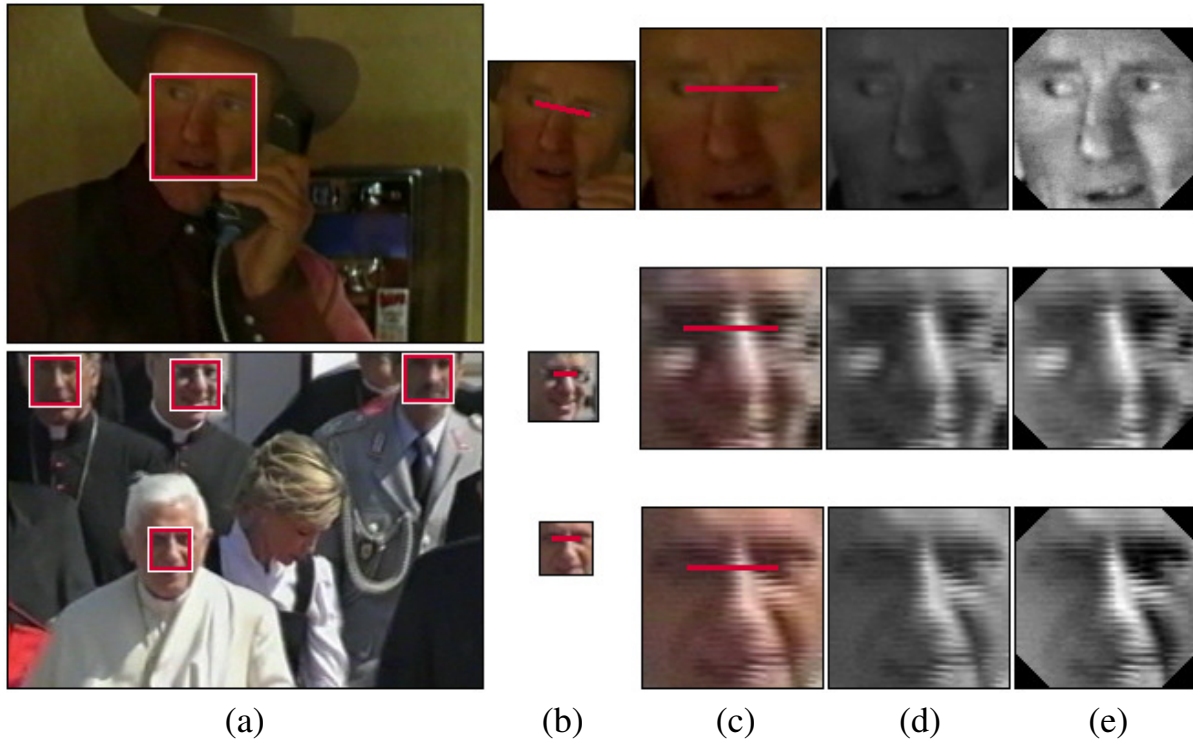


Abbildung 7.5: Segmentierung und Normalisierung von Gesichtsregionen: Automatisch erkannte Gesichtsregionen (a), Lokalisierung der Augen (b), Ausgleich der Rotation und Skalierung des Bildes (c), Umwandlung in ein Graustufenbild (d), Ausgleich von Beleuchtungsunterschieden und Anpassung des Kontrastes (e).

quadratischer Bildbereich wird als Gesicht ausgewählt, dessen Seitenlänge dem doppelten Augenabstand entspricht. Der Bereich wird vertikal verschoben, so dass der Abstand der Augen vom oberen Rand einem Drittel der Gesichtsgröße entspricht. Die Segmentierung wird mit einer Skalierung des Bildausschnitts auf eine einheitliche Größe von 100×100 Pixel abgeschlossen. Abbildung 7.5 (c) zeigt Beispiele für einheitlich skalierte Gesichtsbilder.

Die Gesichtserkennung mit Eigengesichtern reagiert empfindlich bei Beleuchtungsunterschieden. Zunächst wird der Lichteinfall aus unterschiedlichen Richtungen durch eine lineare Funktion angenähert und ausgeglichen [488]. Insbesondere bei schlechter Beleuchtung ist zusätzlich eine Anpassung des Kontrastes notwendig. Die Helligkeitswerte der Gesichtsregion $I_{x,y}$ werden so skaliert, dass die Breite des Intervalls $[0, 255]$ ausgenutzt wird:

$$I'_{x,y} = (I_{x,y} - I_{min}) \cdot \frac{255}{I_{max} - I_{min}} \quad (7.1)$$

I_{min} und I_{max} bezeichnen die minimale und maximale Helligkeit innerhalb der ursprünglichen



Abbildung 7.6: Beispiele für Eigengesichter mit den höchsten Eigenwerten. Zur besseren Darstellung sind die Eigengesichter invertiert dargestellt.

Bildregion. Durch die Skalierung liegen die neuen minimalen und maximalen Helligkeitswerte bei 0 bzw. bei 255. Häufig sind einzelne Pixel in Videos stark verrauscht, so dass ein besonders heller oder dunkler Wert die Anpassung des Kontrastes verhindert. Damit einzelne fehlerhafte Pixel möglichst geringe Auswirkungen auf die Skalierung haben, wird die Gesichtsregion vor der Berechnung der Faktoren I_{min} und I_{max} geglättet. Die Anpassung des Kontrastes erfolgt anschließend auf dem ursprünglichen nicht geglätteten Bild. Damit der Bildhintergrund die Klassifikation nicht beeinflusst, wird in den Ecken ein kleiner dreieckiger Bildbereich entfernt. Abbildung 7.5 verdeutlicht die wesentlichen Schritte der Segmentierung und Normalisierung einer Gesichtsregion. Insbesondere in den skalierten Gesichtsbildern des zweiten Videos sind Fehler durch die analoge Aufnahme und anschließende Digitalisierung deutlich sichtbar. Trotz der schlechten Qualität ist eine zuverlässige Gesichtserkennung mit dem im folgenden Abschnitt vorgestellten Verfahren möglich.

7.3.3 Klassifikation eines Gesichtes

Die Erkennung eines Gesichtes verwendet die Methode von Turk et al. [514, 515]. Aus einer Trainingsmenge mit Gesichtern werden Eigenvektoren ermittelt, die wegen ihres gesichtsähnlichen Aussehens als *Eigengesichter* bezeichnet werden. Die Eigenvektoren spannen als Basisvektoren den sogenannten *Gesichtsraum* auf. Abbildung 7.6 zeigt Beispiele für Eigengesichter mit den größten Eigenwerten.

Die Eigenvektoren mit den größten Eigenwerten beschreiben die wesentlichen Merkmale aller Gesichter der Trainingsmenge, so dass es ausreicht, diese zur Erkennung von Gesichtern zu verwenden [287]. In der Untersuchung einer Trainingsmenge mit 200 Gesichtern decken die ersten 10 Eigenvektoren mehr als 82 Prozent der Varianz der Gesichtsbilder ab, die ersten

50 Eigenvektoren sogar 95 Prozent [269, 461]. Durch die Verwendung der Eigenvektoren mit den größten Eigenwerten wird bei einer Annäherung eines Gesichtes der durchschnittliche quadratische Fehler minimiert.

Die Robustheit der Gesichtserkennung bei Beleuchtungsänderungen, Skalierungen und Rotationen wurde von Turk et al. mit einer umfangreichen Bildsammlung von mehr als 2500 Bildern analysiert [513, 514]. Bei geringen Beleuchtungsunterschieden liegt der Anteil der korrekt erkannten Personen bei 96 Prozent. Deutlich kritischer wirkt sich eine Drehung des Kopfes oder eine Skalierung des Bildes aus, durch die die Erkennungsraten auf 85 Prozent bzw. 64 Prozent absinken. Durch die Segmentierung und Normalisierung der Bilddaten werden Beleuchtungs- und Größenunterschiede zuverlässig ausgeglichen.

7.4 Experimentelle Ergebnisse

Bei der Lokalisierung von Gesichtsregionen mit neuronalen Netzen können abhängig von den analysierten Videosequenzen zwischen 56 und 79 Prozent der frontalen Gesichter gefunden werden [373]. Der Anteil der fehlerhaft als Gesicht klassifizierten Bildbereiche liegt unter 13 Prozent. Durch einen Vergleich der Positionen und Größen der erkannten Gesichtsregionen in benachbarten Bildern können die Fehler deutlich verringert werden. Eine Gesichtsregion gilt nur dann als korrekt lokalisiert, falls innerhalb einer Kameraeinstellung mindestens drei weitere Gesichtsregionen an ähnlicher Position und in vergleichbarer Größe gefunden werden. Einzelne fehlerhafte Regionen werden so erfolgreich ausgefiltert.

Zur Überprüfung der Qualität haben wir Gesichtsregionen in zwei Nachrichtensendungen und zwei Spielfilmen analysiert. Für die Gesichtserkennung mit Eigengesichtern konnten wir in den analysierten Videosequenzen sehr zuverlässige Ergebnisse erreichen [287, 391]. Tabelle 7.1 gibt die Länge der Videos und die Anzahl der erkannten Gesichtsregionen an.

Bei der Suche nach einem Gesicht wird ein Bild der entsprechenden Person in über 90 Prozent der Abfragen korrekt zurückgeliefert. In den analysierten Videos sind vier Ursachen für die Fehler bei der Erkennung mit Eigengesichtern verantwortlich: *Beleuchtungsunterschiede*, eine *seitliche Neigung des Kopfes* (Rotation in der Bildebene), *Skalierungsunterschiede* und eine *Drehung des Kopfes* nach links oder rechts. Da wir die Erkennung der Gesichter auf frontale Gesichter beschränken wollen, spielen lediglich die ersten drei Faktoren eine Rolle. Wird bei der Segmentierung und Normalisierung des Bildes eine Rotation des Kopfes nicht ausgeglichen, so sinkt die Erkennungsrate um fast zehn Prozent. Noch deutlicher wirken sich Beleuchtungs- bzw. Größenunterschiede aus, die den Anteil der korrekt erkannten Gesichter

	Nachrichten 1	Nachrichten 2	Spielfilm 1	Spielfilm 2
Länge	8 min	15 min	142 min	127 min
Anzahl der Bilder	11.587	23.342	204.366	183.504
Anzahl der Gesichtsregionen	4.477 (39%)	10.684 (46%)	47.992 (23%)	31.583 (17%)
Anzahl der Personen	31	47	61	28
Verteilung der Personen	1.Sprecher (14,4%) 2.Sprecher (8,5%) Politiker (5,3%)	Sprecher (29,8%) Politiker (2,9%) Reporter (2,5%)	1.HD (7,9%) 1.HD (5,6%) 2.HD (1,7%)	1.HD (6,1%) 2.HD (3,5%) 3.HD (1,8%)

Tabelle 7.1: *Ergebnisse der Gesichtserkennung: Nachrichtensprecher, Politiker, Reporter und Hauptdarsteller (HD) werden erkannt, wobei in Spielfilm 1 der erste Hauptdarsteller zwei unterschiedlichen Personenklassen zugeordnet wird.*

um 12 bzw. um mehr als 21 Prozent verringern.

Neben der Suche nach einer einzelnen Person innerhalb eines Videos werden weiter gehende semantische Fragestellungen im Rahmen der computergestützten Analyse von Videos untersucht. Gesichtsregionen liefern Informationen über die Anzahl der Personen in einer Kameraeinstellung und ihre Entfernung zur Kamera. Hauptdarsteller bzw. besonders relevante Personen einer Dokumentation können beispielsweise anhand besonders großer Gesichter erkannt werden. Durch Analyse der Position des Gesichtes im Bild kann die Bewegung einer Person innerhalb einer Kameraeinstellung verfolgt werden. Aus der Position eines Gesichtes im Zeitablauf kann in Nachrichtensendungen ein Sprecher oder Reporter erkannt werden.

Die Gesichtserkennung liefert weitere wichtige Informationen, wie beispielsweise die gesamte Dauer, die eine oder mehrere Personen im Video sichtbar sind. Besonders relevante Personen oder Personengruppen des Videos lassen sich so ermitteln. Der Name einer Person kann ausgegeben werden, falls entsprechende Gesichtsbilder in der Trainingsmenge enthalten sind. Für ein Filmarchiv liefert die Suche eines speziellen Gesichtes eine Liste mit Kameraeinstellungen, die auf unterschiedliche Videos verweisen können.

Im Folgenden werden einzelne Fragestellungen untersucht, die semantische Informationen über Personen in Videos liefern. Zwei Spielfilme mit einer Länge von etwas über zwei Stunden und zwei Nachrichtensendungen wurden hierzu analysiert. Fünf Fragestellungen werden exemplarisch betrachtet, deren Ergebnisse in Tabelle 7.1 aggregiert sind:

1. *In wievielen Bildern ist mindestens ein Gesicht abgebildet?*

Der Anteil der Bilder, in denen Gesichtsregionen gefunden werden, liegt in den beiden

Nachrichtensendungen mit 39 bzw. 46 Prozent deutlich höher als in Spielfilmen. Der hohe Anteil in Nachrichtensendungen ist auf die vielen frontalen Gesichtsaufnahmen der Nachrichtensprecher, Reporter und Politiker zurückzuführen. In durchschnittlich 22 Prozent der Gesichtsbilder in den Nachrichtensendungen wird mehr als ein Gesicht gefunden, bei den beiden Spielfilmen liegt der Anteil bei 14 Prozent.

2. *Wieviele unterschiedliche Personen gibt es im Video?*

Alle Gesichter eines Videos werden gespeichert und in den Gesichtsraum transformiert. Zur Gruppierung ähnlicher Gesichter zu einer Gesichtsklasse verwenden wir den *K-Means*-Algorithmus. Nur Gesichtsklassen, die mehr als fünf Gesichter enthalten, werden berücksichtigt, da Personen der kleineren Klassen nur sehr kurz sichtbar sind oder die Gesichtsbilder fehlerhaft oder zum Teil verdeckt sind. Die Anzahl der Gruppen gibt Auskunft, wieviele unterschiedliche Personen im Video vorkommen. Um Personen im Bildhintergrund auszuschließen, haben wir in den beiden Spielfilmen nur große Gesichter mit einer Breite von mindestens zwanzig Prozent der Bildhöhe berücksichtigt.

3. *Welche Personen sind besonders relevant für ein Video?*

Bei dieser Fragestellung wird die Annahme getroffen, dass relevante Personen wie beispielsweise Hauptdarsteller besonders häufig auftreten und in Nahaufnahme gezeigt werden. Die Anzahl der Gesichter in einer Personenklasse liefert den Anteil der Gesichtsbilder dieser Person. Tabelle 7.1 verdeutlicht, dass in den beiden Nachrichtensendungen zwischen 23 und 30 Prozent aller Bilder einen Sprecher zeigen. Wesentlich seltener werden frontale Gesichter eines speziellen Hauptdarstellers erkannt. Um besonders relevante Personen zu finden, wird die Anzahl der erkannten Gesichter einer Personenklasse mit der Gesichtsgröße gewichtet.

4. *Welche Personen treten am häufigsten zu zweit auf?*

Für alle Bilder des Videos mit mindestens zwei erkannten Gesichtsregionen werden die Gesichtsklassen ermittelt. Wir nutzen eine Matrix um zu zählen, wie häufig zwei Gesichtsklassen gleichzeitig in einem Bild auftreten.

5. *In welchen Bildbereichen sind Personen sichtbar?*

Abbildung 7.7 verdeutlicht die Verteilung der Gesichtsregionen im Bild, wobei die Ergebnisse für die beiden Nachrichtensendungen und die Spielfilme zusammengefasst sind. In allen Videos liegt der Schwerpunkt im rechten oberen Bildbereich. Die Regionen des Nachrichtensprechers sind besonders deutlich erkennbar.

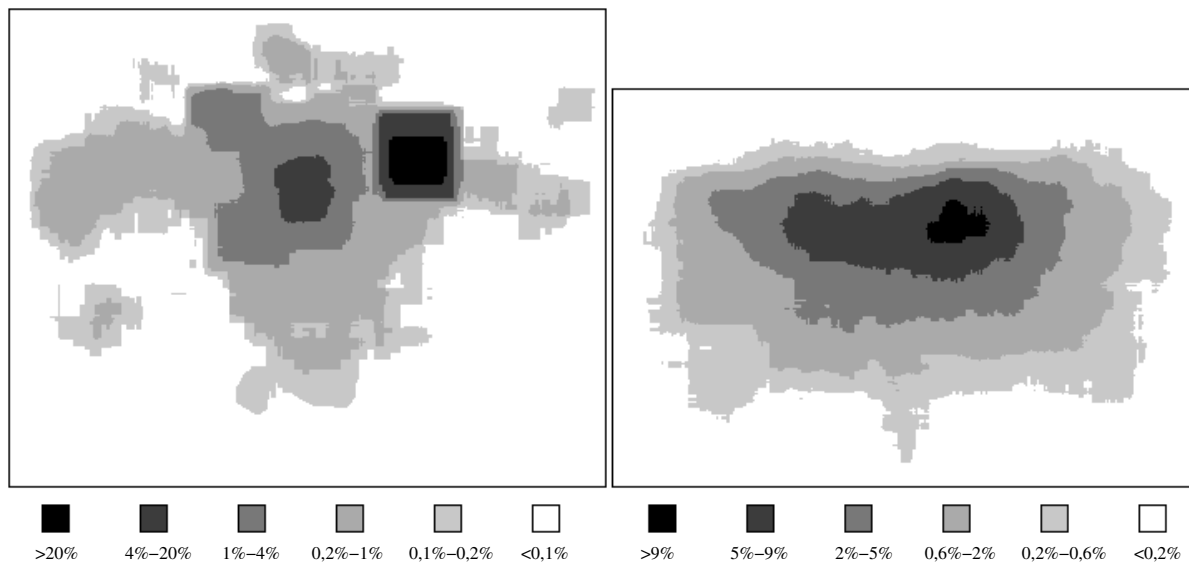


Abbildung 7.7: Verteilung der Gesichter im Bildbereich für eine Nachrichtensendung (links) und einen Spielfilm (rechts).

7.5 Zusammenfassung

In diesem Kapitel wurde zunächst eine Klassifikation bestehender Gesichtserkennungsalgorithmen anhand modellbasierter und konnektionistischer Verfahren eingeführt. Anschließend wurde die Gesichtserkennung als ein dreistufiger Prozess vorgestellt: die Lokalisierung, die Segmentierung (Feinlokalisierung) und Normalisierung sowie die eigentliche Gesichtserkennung. Für den ersten und dritten Schritt wurde auf bekannte Verfahren zurückgegriffen. Im zweiten Schritt wurde ein neuer modellbasierter Algorithmus entwickelt, der eine genaue Segmentierung ermöglicht und zusätzlich Rotationen, sowie Skalierungs-, Kontrast- und Beleuchtungsunterschiede ausgleicht.

In den experimentellen Ergebnissen wurde insbesondere darauf eingegangen, wie neue semantische Informationen aus erkannten Gesichtern abgeleitet werden können. Anhand von fünf untersuchten Fragestellungen wurde deutlich, dass Gesichter wichtige semantische Informationen über ein Video liefern. Diese Informationen sind nicht nur für die Indexierung von Videos relevant, sondern bieten die Möglichkeit, gute Algorithmen zur Adaption von Videos zu entwickeln. Verfahren, die Zusammenfassungen eines Videos automatisch erzeugen, profitieren ganz wesentlich von den Ergebnissen der Objekt-, Text- und Gesichtserkennungsalgorithmen.

Teil II

Anwendungen zur Analyse digitaler Videoarchive

KAPITEL 8

Adaption von Videos

Durch den technologischen Fortschritt der letzten Jahre ist die Wiedergabe eines Videos nicht mehr auf Fernseher oder PCs beschränkt, sondern auf einer Vielzahl von Geräten möglich, die hinsichtlich ihrer Ausstattungsmerkmale deutlich variieren. Insbesondere die Größe der Displays und die unterschiedlichen Übertragungskapazitäten der Netzwerke führen dazu, dass Videos auf vielen Geräten nur mit deutlichen Einschränkungen betrachtet werden können.

Eine besonders starke Verringerung der Qualität ist häufig bei der Wiedergabe eines Videos auf einem mobilen Gerät zu beobachten. Obwohl aktuelle mobile Geräte über ausreichende Rechenkapazitäten verfügen, müssen noch grundlegende Probleme gelöst werden, bis bestehendes Videomaterial auf diesen Geräten in guter Qualität wiedergegeben werden kann. Eine große Herausforderung ist die *Heterogenität der unterschiedlichen Geräte*. Neben der Einteilung in Geräteklassen wie beispielsweise Notebooks, Tablet-PCs, Handheld-PCs (PDAs) oder Mobiltelefone differieren die einzelnen Geräte auch deutlich innerhalb ihrer Klasse. Zu den wesentlichen Eigenschaften zählen die Auflösung und Farbtiefe des Displays, die Größe des Arbeitsspeichers, die Leistungsfähigkeit des Prozessors und die verfügbare Software zur Dekodierung und Darstellung eines Videos. Aufgrund der beschränkten Speicherkapazität werden Videos im Allgemeinen erst beim Abspielen auf das mobile Gerät übertragen, so dass auch die Übertragungskapazität der in das Gerät integrierten Kommunikationsschnittstelle zum Engpass werden kann.

Für eine gute Darstellung sollten bestehende Videos möglichst genau an die unterschiedlichen Eigenschaften der Anzeigegeräte angepasst werden. Eine manuelle Festlegung der Parameter für alle Kombinationen von Videos und Anzeigegeräten verursacht durch die deutlichen

Unterschiede bezüglich der Hardware, der Software und der verfügbaren Netzwerkkapazität einen sehr hohen Aufwand. Verfahren zur *automatischen Adaption von Videos* ermöglichen die Wiedergabe bestehender Videos auch auf mobilen Geräten ohne zusätzlichen Aufwand. Das zentrale Ziel der Adaption ist der *Erhalt der semantischen Informationen* eines Videos unabhängig von der Ausstattung eines Anzeigerätes.

Die wesentlichen Parameter eines Videos, die bei der Adaption geändert werden müssen, sind die Bitrate, die Farbtiefe, die Bildauflösung und die Bildwiederholrate, wobei die letzten beiden Parameter die Bitrate wesentlich beeinflussen. Zur Anpassung der Bildauflösung ist eine Skalierung des Bildes nicht optimal, falls Bildinhalte wegen ihrer geringen Größe nicht oder nur noch sehr schwer erkannt werden können. Ein intelligentes Verfahren zur Anpassung der Bildgröße, das semantische Inhalte eines Videos berücksichtigt, kann die Qualität des adaptierten Videos deutlich erhöhen.

Auch die Qualität des ursprünglichen Videos entspricht nicht immer den Erwartungen eines Betrachters. Eine Verbesserung der Bildqualität des ursprünglichen Videos wirkt sich auch auf das adaptierte Video aus, da deutlich mehr Details erkannt werden können. Insbesondere in Amateurvideos oder historischen Filmen sind durch die Lagerung der Bänder und die mangelhafte Aufzeichnungstechnik der Kameras viele Bildfehler im Video enthalten, die bei der Adaption ausgeglichen werden sollten. Zu den typischen Fehlern zählen über- oder unterbelichtete Kameraeinstellungen, Helligkeitsschwankungen, eine verwackelte Kameraführung oder Streifen und Kratzer im Bild.

In diesem Kapitel wird zunächst ein Überblick über Verfahren zur Adaption von Videos gegeben, wobei viele bestehende Ansätze lediglich eine effiziente Anpassung der formatspezifischen Parameter eines Videos ermöglichen. In den folgenden Abschnitten werden neue Verfahren zur semantischen Adaption eines Videos vorgestellt. Die *Adaption der Farbtiefe* zur Darstellung eines Videos auf einem Graustufendisplay erfolgt durch Analyse der Helligkeitsverteilung der Pixel einer Kameraeinstellung. Ein zweites neues Adaptionsverfahren zur Erzeugung von Binärbildern kombiniert Kanteninformationen mit Texturen [284].

Die *Adaption der Bildauflösung* ist durch das Abschneiden der Bildränder oder eine Skalierung des Bildes möglich. Ein neues Verfahren wird vorgestellt, das semantische Inhalte des Videos analysiert, bewertet, zu Regionen zusammenfasst und die Region mit der höchsten Bewertung für das adaptierte Video auswählt [286]. Bei mehreren gleichwertigen Regionen wird ein künstlicher Kameraschwenk zwischen diesen Regionen erzeugt.

Ein drittes neues Adaptionsverfahren wird zur *Verbesserung der Bildqualität* von Amateurvideos und historischen Videos vorgeschlagen, so dass Bildinhalte im adaptierten Video besser

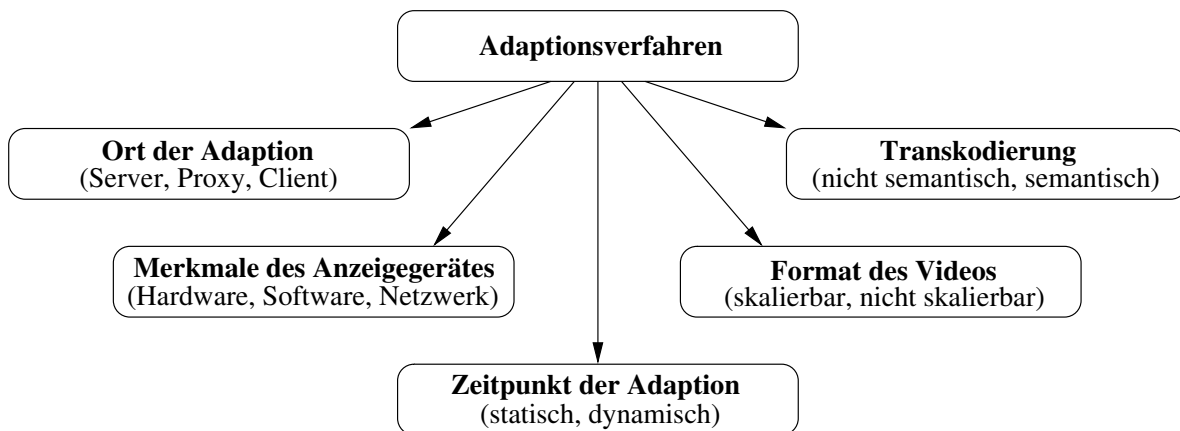


Abbildung 8.1: Klassifikation der Verfahren zur Adaption von Videos

erkannt werden können und das Betrachten des Videos angenehmer wird. Bei diesem Verfahren werden die Helligkeit und der Kontrast angepasst, Streifen und Kratzer im Bild entfernt und verwackelte Aufnahmen stabilisiert. Experimentelle Ergebnisse zu den entwickelten Adaptionsverfahren schließen das Kapitel ab.

8.1 Verfahren zur Adaption multimedialer Inhalte

Die *Adaption von multimedialen Inhalten* (engl. *content repurposing*) soll die Wiedergabe auf Geräten mit unterschiedlichen Ausstattungsmerkmalen in guter Qualität ermöglichen. Die bestehenden Verfahren zur Adaption von Videos können anhand unterschiedlicher Charakteristiken entsprechend Abbildung 8.1 klassifiziert werden [311]. Die Adaption wird auf einem *Server* [359, 208, 387], einem *Proxy* [186, 335] oder direkt auf dem *Client* [301] durchgeführt. Bei einer serverbasierten Lösung wird es insbesondere bei einer großen Anzahl von Clients durch den für die Adaption eines Videos erforderlichen Rechenaufwand zu Performanceengpässen kommen. Auf der anderen Seite stehen clientbasierte Ansätze, die für Videos wegen der großen Datenmengen im Allgemeinen nicht geeignet sind [72].

Die *technischen Merkmale* eines Gerätes in Form von Hardware, Software und der aktuell verfügbaren Netzwerkkapazität definieren die formatspezifischen Parameter der Adaption [147]. Ein weiteres Klassifikationskriterium betrifft den *Zeitpunkt der Adaption*. Abhängig von der Komplexität der Adaptionsalgorithmen kann eine Berechnung in Echtzeit nicht immer gewährleistet werden, so dass im Vorfeld mehrere statische Versionen eines Videos für ausgewählte Geräteprofile berechnet und gespeichert werden müssen. Bei einer dynamischen Adaption er-

folgt die Berechnung und Transkodierung des Videos in Echtzeit.

Falls das *Format des Videos* eine Skalierung unterstützt, können in einem Videostrom mehrere Versionen in unterschiedlichen Qualitätsstufen kodiert sein. Die Basisschicht (engl. *base layer*) speichert das Video mit sehr geringer Qualität und benötigt die wenigsten Ressourcen für die Darstellung. Bei zusätzlich verfügbaren Kapazitäten werden weitere Schichten (engl. *enhancement layer*) zur Verbesserung der Qualität des Videos übermittelt.

Damit das Video dargestellt werden kann, muss der Adaptionalgorithmus die Merkmale des Anzeigegeätes, also die Farbtiefe, die Bildauflösung, die Bildwiederholrate und die Bitrate, berücksichtigen. Die Anpassung der Parameter erfolgt bei der *Transkodierung* des Videos. Bei einer *semantischen Transkodierung* sollen die Bildinhalte des Videos analysiert und geeignete Parameter für den Adaptionalgorithmus so spezifiziert werden, dass wichtige Bildinhalte nach der Adaption möglichst gut erkannt werden können [390].

8.1.1 Unterstützung der Adaption durch Standardisierungsverfahren

Die beiden Standards *MPEG-7* und *MPEG-21* unterstützen die semantische Beschreibung der Inhalte eines Videos [228, 230]. Beide Formate ermöglichen es, Informationen zur Personalisierung und Adaption eines Videos zu speichern [508, 526]. *MPEG-7* umfasst eine Datenbeschreibungssprache zum vereinfachten Austausch multimedialer Daten. Zusätzlich wird der netzbasierte Zugriff von beliebigen Geräen auf multimediale Daten unterstützt, der unter dem Begriff *Universal Multimedia Access* zusammengefasst wird [28, 359, 528]. Regeln für die Transkodierung von Videos, eine Nutzerhistorie und individuelle Nutzerpräferenzen (engl. *user preference description*) können gespeichert werden, die zusätzliche Informationen für den Adaptionalgorithmus bereitstellen.

MPEG-21 erweitert die verfügbaren Metadaten und ermöglicht eine Beschreibung der *Gerätemerkmale* (engl. *usage environment description*). Innerhalb dieser Beschreibung sind Daten zur Charakterisierung des Displays, der Systemkonfiguration sowie der verfügbaren Hardware und Software vorgesehen. Zusätzliche Techniken, wie beispielsweise die Modellierung von Nutzeranfragen und Nutzerpräferenzen, sind im Rahmen von *MPEG-21* standardisiert [388]. Für alle *digitalen Elemente* (engl. *digital item*) innerhalb von *MPEG-21* können spezielle Adaptionsverfahren definiert werden (engl. *digital item adaptation*) [229].

8.1.2 Verfahren zur Adaption von Videos

Um einen Überblick über Algorithmen zur *Adaption von Videos* zu geben, werden zunächst Verfahren zur Adaption von Bildern und Audiodateien betrachtet. Bei der *Bildadaption* ist eine Anpassung an die physikalischen Merkmale des Displays erforderlich, also die Verringerung der Farbtiefe und der Bildauflösung [253, 431]. Jede Adaption eines Bildes sollte das Ziel verfolgen, die Bildinhalte verständlich und vollständig darzustellen.

Bei einer Verringerung der Farbtiefe können wichtige semantische Informationen verloren gehen. Verstärkt tritt dieses Problem bei der Darstellung von Bildern auf Schwarz-Weiß-Displays auf [430]. Eine Verkleinerung des Bildes liefert akzeptable Ergebnisse nur bis zu einem gewissen Grad, da der Inhalt mit zunehmender Skalierung immer schwieriger erkannt wird. Die Anpassung der Bildgröße ist durch das Abschneiden von Rändern oder die Verkleinerung des Bildes möglich. Die Auswahl der Bildregion sollte so erfolgen, dass Objekte, die die Aufmerksamkeit eines Betrachters auf sich ziehen (engl. *attention object*), auch nach der Adaption noch erkannt werden können [78, 133]. Diese Objekte können zu Regionen mit wichtigen semantischen Informationen zusammengefasst werden (engl. *region of interest*) [210, 434].

Anhand der Farbverteilung, des Kontrastes und der Orientierung der Kanten im Bild können wichtige Bildregionen identifiziert werden [231]. Bei sehr großen Bildern mit vielen detaillierten Informationen, wie beispielsweise einer technischen Zeichnung, können durch das Abschneiden der Bildränder oder eine Skalierung des Bildes sehr viele wichtige Informationen verloren gehen. In diesen Fällen bietet sich eine Unterteilung des Bildes in mehrere kleine Bilder [253] oder die Umwandlung in eine Animation oder ein Video an, bei dem ein künstlicher Kameraschwenk die unterschiedlichen Bildausschnitte hervorhebt [327].

Bei der *Adaption eines Audiosignals* werden zwei unterschiedliche Ansätze betrachtet. Zunächst können die formatspezifischen Merkmale des Audiosignals in Form von Frequenzumfang oder der Art der Kodierung angepasst werden. Durch eine Beschleunigung der Abspielgeschwindigkeit wird die zeitliche Länge des Audiosignals reduziert. Dabei sollte die Tonhöhe unverändert bleiben, damit die Sprache verständlich bleibt [524]. Die zweite Gruppe der Verfahren wandelt das Audiosignal in eine andere Darstellungsform um. Dabei ist insbesondere die Spracherkennung, also die Umwandlung des akustischen Signals in einen Text, wichtig, wie sie beispielsweise zur Indexierung von Nachrichtensendungen eingesetzt wird [160, 474].

Wegen der großen Datenmenge sollten bei der *Adaption eines Videos* effiziente Verfahren zur *Transkodierung* eingesetzt werden [527]. Der Wechsel des Kompressionsverfahrens ist erforderlich, falls ein Video in einem speziellen Format wegen unzureichender Hardware oder feh-

lender Software nicht abgespielt werden kann [20, 38, 104, 302]. Zusätzlich werden bei der Transkodierung eines Videos die formatspezifischen Parameter in Form von Bitrate, Bildauflösung, Farbtiefe oder Bildwiederholrate mit möglichst geringem Rechenaufwand angepasst [458]. Zur Verringerung der Rechenzeit werden Ergebnisse von Berechnungen aus dem ursprünglichen Video, wie beispielsweise die Ermittlung der Bewegungsvektoren, wiederverwendet.

Durch die Analyse der semantischen Inhalte eines Videos können einzelne Bildregionen im adaptierten Video hervorgehoben werden [32, 475, 570]. Objekte und Ereignisse liefern Informationen über wichtige Bildregionen innerhalb einer Kameraeinstellung [259, 507, 509, 529]. Mehrere Systeme zur automatischen Adaption von Videos sind in Forschungsergebnissen beschrieben, wobei viele bestehende Verfahren ihren Schwerpunkt auf die effiziente Transkodierung eines Videos legen [204, 379, 475].

Falls nur eine geringe Netzkapazität zur Verfügung steht, ist die Übertragung eines Videos in Echtzeit nicht möglich. In diesen Fällen bietet sich die Darstellung des Videos als Folge von einzelnen aussagekräftigen Bildern (engl. *key frame*) an [571]. Bei unzuverlässigen Netzverbindungen müssen Teile oder das gesamte Video vor der Wiedergabe auf das Anzeigegerät übertragen werden, wobei die Speicherung eines längeren Videos beim Empfänger wegen der großen Datenmenge nicht immer möglich ist. Zusammenfassungen von Videos (engl. *video summary*), auf die detailliert in Kapitel 9 eingegangen wird, bieten auch bei eingeschränkten Netzverbindungen die Möglichkeit, die wesentlichen Inhalte eines Videos in kompakter Form wiederzugeben [167, 463, 483].

Adaptionsmöglichkeiten bestehen nicht nur in der Anpassung eines Videos an die Merkmale eines Anzeigegerätes, sondern auch in der Qualitätsverbesserung eines Videos. In Amateurvideos und historischen Filmen ist der Anteil fehlerhafter Kameraeinstellungen besonders hoch [113, 272, 444]. Bildfehler können durch natürliche Alterung der Filme, eine Verschmutzung der Filmrolle [23, 455] oder durch Abnutzung beim Transport der Filmrolle im Projektor entstehen [52, 244, 243]. Als besonders störend werden horizontale oder vertikale Linien im Bild empfunden [51, 271]. Falsch belichtete Kameraeinstellungen oder verwackelte Aufnahmen sind weitere häufig zu beobachtende Fehler [273, 561]. Bei einer geringen Qualität des ursprünglichen Videos können die Bildinhalte im adaptierten Video häufig nicht erkannt werden. Eine Lösung bieten Verfahren zur Verbesserung der Bildqualität eines Videos.

8.2 Anpassung der Farbtiefe eines Videos

Bei einer Verringerung der Farbtiefe auf wenige Helligkeitswerte können große Regionen mit gleichen Helligkeitswerten entstehen, so dass der Bildinhalt in Teilen des Videos nicht mehr erkannt werden kann. Eine besondere Herausforderung liegt in der Adaption eines Videos für monochrome Displays, in denen die Bilder durch zwei unterschiedliche Helligkeitswerte dargestellt werden.

Die Umwandlung der Farbpixel in Graustufenwerte ist in Videos ohne zusätzlichen Rechenaufwand möglich, da die Helligkeit unabhängig von den Farbinformationen ähnlich dem YUV-Farbmodell gespeichert wird. Bei einer Verringerung der Anzahl der unterschiedlichen Helligkeitswerte gehen Details des Bildes verloren, was zunächst bei fließenden Übergängen, Helligkeitsverläufen und feinen Texturen zu deutlich wahrnehmbaren Fehlern führt. Bei der Analyse des Histogramms eines Bildes wird deutlich, dass die Verteilung der Pixel in der Regel nicht gleichmäßig ist und in einem großen Anteil der analysierten Testbilder viele Pixel innerhalb weniger Intervalle liegen. Bei einer linearen Adaption der Helligkeit werden Intervalle gleicher Größe definiert. Da alle Werte innerhalb eines Intervalls die gleiche Helligkeit zugewiesen bekommen, wird der Kontrastumfang des Displays nicht ausgeschöpft und viele Details gehen verloren.

Zunächst schlagen wir ein Verfahren zur Verringerung der Farbtiefe vor und erläutern es am Beispiel die Adaption von 256 auf 8 unterschiedliche Helligkeitswerte. Bei Graustufenbildern mit 256 unterschiedlichen Helligkeiten werden bei einer *linearen Adaption* Intervalle gleicher Größe definiert. Durch die Anzahl der Helligkeitswerte N_C im adaptierten Bild wird die Intervallgröße $\frac{256}{N_C}$ bestimmt. Alle Helligkeitswerte innerhalb eines Intervalls werden auf einen neuen Helligkeitswert abgebildet:

$$I_{lin}(i) = \lfloor \frac{N_C}{256} \cdot i \rfloor \in [0, N_C - 1]. \quad (8.1)$$

Die Pixel des ursprünglichen Bildes mit der Helligkeit i erhalten durch die Adaption den neuen Wert $I_{lin}(i)$ zugewiesen. Abbildung 8.2 (c) verdeutlicht, dass bei einer Verringerung der Farbtiefe auf acht Helligkeitswerte insbesondere feine Strukturen verloren gehen. Eine variable Größe der Intervalle abhängig von der Verteilung der Helligkeitswerte liefert mehr Detailinformationen im adaptierten Bild, insbesondere für Bilder mit einem geringen Kontrast. Eine nicht lineare Abbildung der Helligkeitswerte ist durch kumulierte Histogramme $H_{kum}(i)$ möglich:

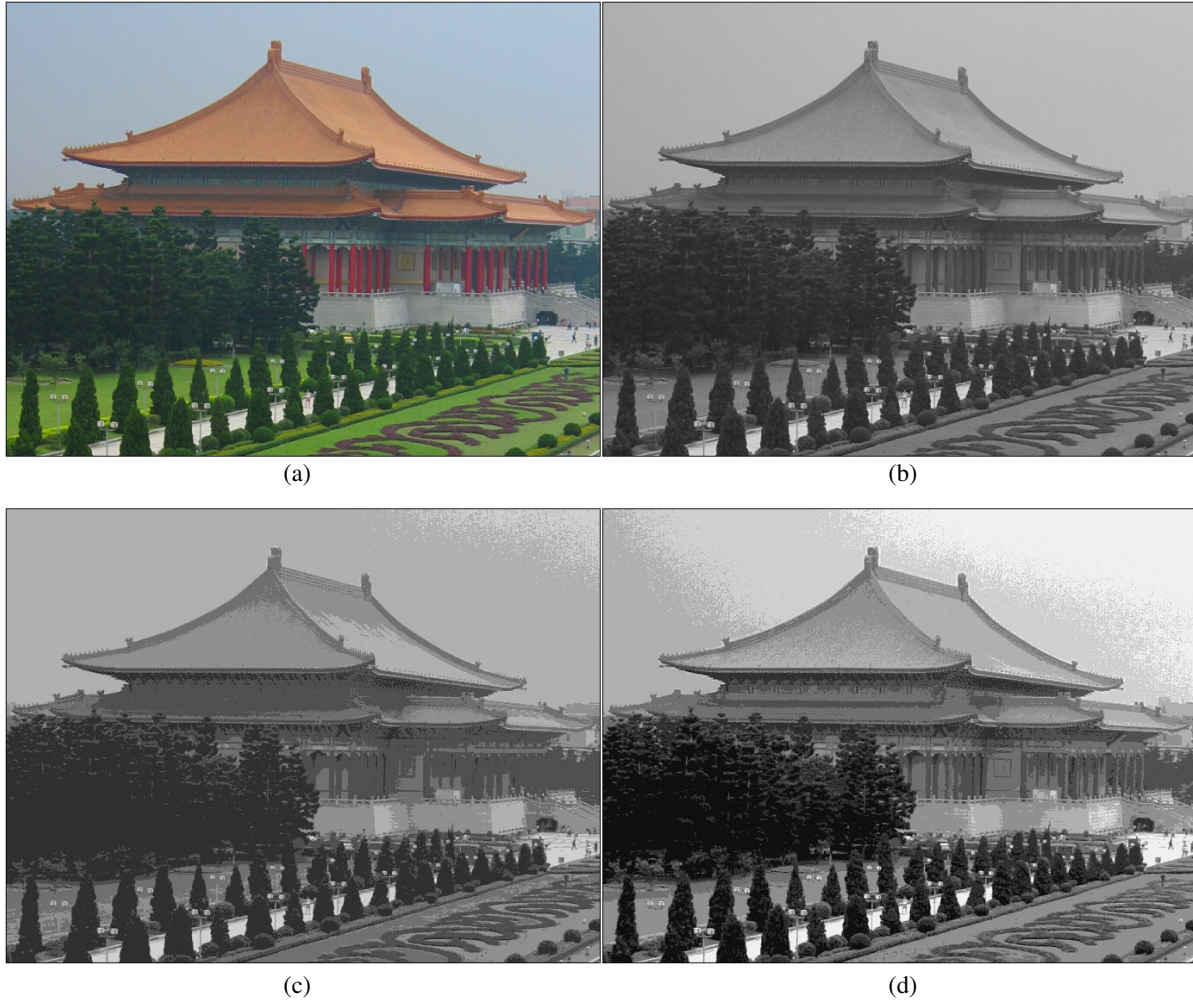


Abbildung 8.2: Transformation eines Farbbildes (a) in ein Graustufenbild mit 256 Helligkeitswerten (b) sowie 8 Helligkeitswerten bei linearer (c) und adaptiver Transformation (d).

$$I_{var}(i) = \lfloor \frac{N_C}{S_X \cdot S_Y + 1} \cdot H_{kum}(i) \rfloor \in [0, N_C - 1]. \quad (8.2)$$

Die Bildbreite S_X und die Bildhöhe S_Y skalieren die Werte des kumulierten Histogramms. Ein Helligkeitswert i wird in Abhängigkeit von der Verteilung der Pixel im kumulierten Histogramm auf den neuen Helligkeitswert $I_{var}(i)$ abgebildet. Die Abbildung 8.2 (d) verdeutlicht, dass durch variable Intervallgrößen mehr Details im Bild erkannt werden können. Bei sehr hellen oder sehr dunklen Bildern führt die nichtlineare Adaption zu einer deutlichen Veränderung der durchschnittlichen Helligkeit. Eine maximal zulässige Änderung der durchschnittlichen Helligkeit kann durch eine Kombination beider Verfahren garantiert werden:

$$L_w(i) = \lfloor \alpha \cdot L_{lin}(i) + (1 - \alpha) \cdot L_{var}(i) \rfloor \in [0, N_C - 1]. \quad (8.3)$$

Der Faktor $\alpha \in [0, 1]$ legt die Gewichtung des linear adaptierten Bildes fest.

In einem weiteren Schritt wird die *Adaption in ein Binärbild* mit nur zwei unterschiedlichen Helligkeiten betrachtet. Das Problem der Darstellung eines Bildes mit einer stark begrenzten Anzahl von Farben oder Helligkeitswerten ist ein bekanntes Problem aus der Drucktechnik. Falls eine Druckmaschine nur wenige Farben wiedergeben kann (keine Halbtöne), wird die Technik als *Offsetdruck* bezeichnet. Dabei werden Bilder gerastert und als feine Punkte nebeneinander bzw. übereinander gedruckt. Um die Druckfarben zu erhalten, werden die Farben eines Bildes auf die neue Farbpalette abgebildet. Im Fall von Binärbildern ist eine Zuordnung der Pixel durch den Vergleich mit einem Schwellwert möglich. Die Abbildungen 8.3 (a) und (b) verdeutlichen am Beispiel von zwei unterschiedlichen Schwellwerten, dass viele Detailinformationen im Bild verloren gehen können.

Beim Offsetdruck erfolgt die Variation der Farb- bzw. Helligkeitswerte durch die Größe der Rasterpunkte (*amplitudenmodulierte Raster*) oder die Anzahl der Punkte pro Fläche (*frequenzmodulierte Raster*). Durch eine geeignete Anordnung der Farbwerte nimmt das menschliche Auge die einzelnen Pixel als gemischte Farbe wahr, so dass die Farbtiefe des adaptierten Bildes deutlich höher zu sein scheint.

Der 1975 veröffentlichte *Floyd-Steinberg-Algorithmus* versucht, den für das menschliche Auge sichtbaren Fehler bei einer Verringerung der Farbtiefe eines Bildes möglichst gering zu halten [145]. Das Bild wird pixelweise von links oben nach rechts unten umgewandelt, wobei das aktuelle Pixel auf die ähnlichste verfügbare Farbe oder Helligkeit abgebildet wird. Der durch den neuen Wert des Pixels entstandene Fehler wird auf benachbarte Pixel verteilt (engl. *error diffusion*). $\frac{7}{16}$ des Fehlers wird auf das rechte benachbarte Pixel, jeweils $\frac{3}{16}$, $\frac{5}{16}$ und $\frac{1}{16}$ auf die angrenzenden Pixel in der folgenden Zeile übertragen. Abbildung 8.3 (c) verdeutlicht, dass der Floyd-Steinberg-Algorithmus im Vergleich zur Umwandlung durch den Vergleich mit einem Schwellwert zu deutlich besseren Ergebnissen führt. Für Videosequenzen ist der Algorithmus jedoch *nicht anwendbar*, da sich durch die Verteilung des Fehlers viele Pixel in aufeinander folgenden Bildern ändern. Das führt zu sehr starkem Rauschen, so dass die Inhalte einer Videosequenz bei der Adaption mit dem Floyd-Steinberg-Algorithmus in sehr schlechter Qualität dargestellt werden.

Eine alternative Darstellung eines Bildes im Binärformat ist durch Kantenbilder möglich. Im Kantenbild der Abbildung 8.3 (d) sind zwar viele Details enthalten, zusammenhängende Flä-

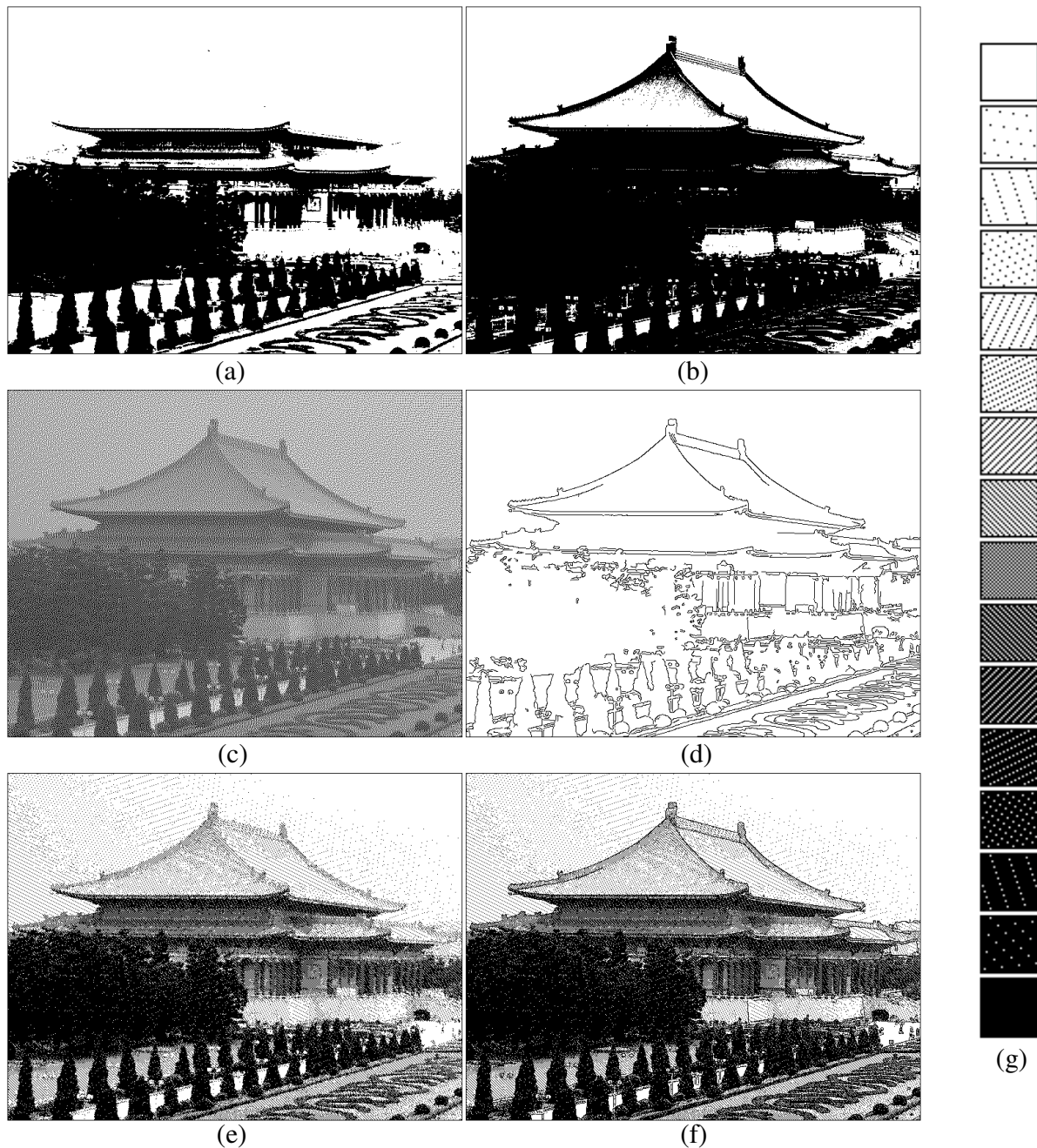


Abbildung 8.3: Transformation eines Farbbildes in ein Binärbild durch Vergleich mit einem Schwellwert von 90 (a) bzw. 130 (b). Der Floyd-Steinberg-Algorithmus (c) liefert gute Ergebnisse, ist jedoch für Videosequenzen nicht anwendbar. Obwohl feine Strukturen im Kantenbild (d) erhalten bleiben, können zusammenhängende Regionen nur schwer erkannt werden. Durch den Einsatz von Texturen (g) können im adaptierten Bild (e) deutlich mehr Inhalte erkannt werden. Die zusätzliche Überlagerung mit Kanten (f) führt zu sehr guten Ergebnissen bei der Adaption von Videos.

chen und Strukturen können jedoch nicht erkannt werden.

Um Bilder mit mehr Details zu erzeugen, die auch in Videosequenzen zu einer guten Darstellung führen, werden 16 binäre Texturen $I_{\text{Textur}}(x, y)$ definiert, die die Pixel im Graustufenbild ersetzen:

$$I_{\text{Textur}}(x, y) = \begin{cases} 0 & [(x + S_X \cdot y) \bmod (T_B + T_W)] < T_B, \\ 1 & \text{sonst.} \end{cases} \quad (8.4)$$

Die beiden Werte T_B und T_W definieren das Verhältnis der schwarzen zu den weißen Pixeln einer Textur. Der Wert von T_B liegt zunächst deutlich über T_W , wobei sich mit jeder weiteren Textur das Verhältnis in Richtung heller Pixel verschiebt. Abbildung 8.3 (g) verdeutlicht die Texturen, die mit Hilfe der Gleichung 8.4 berechnet werden. Die Werte für T_B und T_W wurden so gewählt, dass sich die Muster der Texturen mit ähnlicher Helligkeit deutlich voneinander unterscheiden. Zusammenhängende Flächen werden dadurch leichter erkannt.

Um ein Bild in ein texturiertes Binärbild umzuwandeln, wird ein Graustufenbild mit $N_C = 16$ unterschiedlichen Helligkeiten erzeugt und jeder Helligkeitswert durch ein Pixel der entsprechenden Textur ersetzt. Trotz der unterschiedlichen Muster der Texturen erscheinen die Übergänge zwischen benachbarten Regionen fließend. Obwohl dieser Effekt bei langsamen Farbverläufen wie beispielsweise dem Himmel in Abbildung 8.3 (e) zu guten Ergebnissen führt, verschwinden auch starke Kanten des Bildes. Alle Kantenpixel des Kantenbildes werden daher in das texturierte Binärbild übernommen. Im Vergleich zu den anderen Binärbildern sind in der Abbildung 8.3 (f) deutlich mehr Bildinhalte erkennbar.

Eine Erweiterung zur *Adaption der Farbtiefe eines Videos* wird im folgenden Schritt betrachtet. Für die Darstellung eines Videos ist es besonders wichtig, dass die Parameter innerhalb einer Kameraeinstellung unverändert bleiben, da sonst deutliche Helligkeitsschwankungen zwischen benachbarten Bildern entstehen. Statt das kumulierte Histogramm für ein einzelnes Bild zu berechnen, werden alle Bilder einer Kameraeinstellung gleichzeitig analysiert. Das kumulierte Histogramm aller Bilder beschreibt die Verteilung der Helligkeitswerte der Kameraeinstellung und liefert einheitliche Parameter $L_{\text{var}}(i)$ zur Adaption der Bilder dieser Kameraeinstellung. Um Verzerrungen des Histogramms durch einzelne sehr helle bzw. sehr dunkle Bilder innerhalb einer Kameraeinstellung zu vermeiden, wie sie beispielsweise bei Blitzlicht oder einer Ausblendung zu beobachten sind, bleiben diese Bilder bei der Berechnung des kumulierten Histogramms unberücksichtigt. Falls eine Umwandlung in ein Binärbild erforderlich

ist, werden die 16 Helligkeitswerte durch entsprechende Texturpixel ersetzt.

8.3 Anpassung der Bildauflösung eines Videos

Neben der Farbtiefe hat auch die Bildauflösung des Displays einen wesentlichen Einfluss auf die Darstellung eines Videos. Der Adaptionalgorithmus muss gewährleisten, dass wichtige Bildinhalte auch bei einer deutlichen Verkleinerung der Bildauflösung erkannt werden können [275]. Die Anpassung der Bildgröße eines Videos wird mittels Skalierung oder durch eine Auswahl einer Bildregion erreicht, bei der die Bereiche außerhalb der Region unberücksichtigt bleiben. Durch eine Veränderung der ausgewählten Bildregionen im Zeitablauf entstehen künstliche Schnitte, Kameraschwenks oder Zoomoperationen, die einzelne Bildinhalte des Videos hervorheben. So ist es beispielsweise möglich, in einem adaptierten Video zu Beginn einer Kameraeinstellung das gesamte Bild zu zeigen und anschließend auf ein einzelnes Objekt zu zoomen. Auch bei einer geringen Auflösung des Displays werden durch die künstlichen Kamerabewegungen sowohl allgemeine Informationen des Bildhintergrundes als auch Details über ein Objekt wiedergegeben.

Um die Bildauflösung eines Videos zu reduzieren, ist eine *Skalierung* (engl. *scaling*) oder ein *Abschneiden von Rändern* (engl. *cropping*) möglich. Durch die Kombination der beiden Verfahren werden wichtige Regionen innerhalb einer Kameraeinstellung hervorgehoben [286]. Vier Heuristiken werden eingesetzt, um Bildregionen in einem Video auszuwählen [280]:

- Regionen, in denen semantisch wichtige Inhalte erkannt werden, sollen im adaptierten Video enthalten sein. Falls ein semantisches Merkmal durch die Verkleinerung des Bildes nicht mehr erkannt werden kann, sollte ein anderer aussagekräftigerer Bildbereich gewählt werden. Bei der Analyse des Videos werden Textregionen, Gesichter und Objekte als semantisch wichtig identifiziert und berücksichtigt.
- Regionen ohne aussagekräftigen Bildinhalt sollen nicht im adaptierten Video enthalten sein. Hierzu zählen der dunkle Randbereich eines Videobildes oder große einfarbige Flächen, die an den Bildrand angrenzen.
- Eine ausgewählte Bildregion wird auf die gewünschte Bildgröße des Videos skaliert. Um Verzerrungen zu vermeiden, sollte das Seitenverhältnis der ausgewählten Region mit dem des adaptierten Videos übereinstimmen.

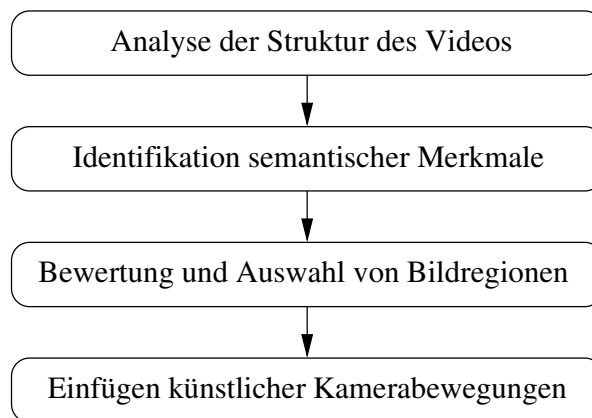


Abbildung 8.4: *Adaption der Bildauflösung eines Videos*

- In Videos ist es möglich, durch einen Wechsel des Bildausschnittes in einer Kameraeinstellung sowohl einen Überblick zu geben als auch Detailinformationen darzustellen. Hierzu werden innerhalb des größeren Originalvideobildes künstliche Kameraraschwenks, Zoomoperationen oder Schnitte mit kleinerem Bildausschnitt eingefügt.

In der Abbildung 8.4 sind die wesentlichen Schritte bei der Adaption der Bildauflösung eines Videos dargestellt. Nach der Analyse der Struktur des Videos werden semantische Merkmale wie beispielsweise Gesichter, Objekte oder Textregionen identifiziert und durch rechteckige Bereiche beschrieben. In einem dritten Schritt wird für jedes einzelne Bild einer Kameraeinstellung eine Bildregion festgelegt, so dass nach der Skalierung dieser Region die Menge der dargestellten Informationen maximal ist. Bei mindestens zwei semantischen Merkmalen im Bild werden in Kameraeinstellungen mit einer gewissen Länge künstliche Kameraoperationen eingefügt. Nach der Festlegung der Bildregionen werden sie passend skaliert und als Video gespeichert. Die Audiospur bleibt bei der Adaption der Bildgröße unverändert.

8.3.1 Identifikation der semantischen Merkmale in Videos

Zur Festlegung der Bildregion des adaptierten Videos werden ausgewählte semantische Inhalte eines Bildes berücksichtigt. Erweiterungen sind möglich, indem beispielsweise Regionen mit starkem Kontrast oder auffälligen Farben identifiziert werden. Es wird angenommen, dass *Gesichtsregionen*, die durch quadratische Bildbereiche beschrieben werden, von zentraler Bedeutung für das Verständnis eines Videos sind. Diese sollen vollständig und möglichst groß im skalierten Video sichtbar sein. Das in Kapitel 7 vorgestellte Verfahren wird zur Erkennung der frontalen Gesichter im Video eingesetzt.

Texte liefern nur dann zusätzliche Informationen über ein Video, wenn eine *Textregion* vollständig und in einer akzeptablen Größe im skalierten Video dargestellt ist. Textregionen werden durch rechteckige Bereiche beschrieben und mit dem Verfahren aus Kapitel 6 identifiziert. *Objektregionen* beschreiben zum Beispiel Personen oder Fahrzeuge im Bildvordergrund, die sich relativ zum Bildhintergrund bewegen. Nach der Segmentierung der Objekte erfolgt die Erkennung durch Analyse der Skalenraumabbildungen entsprechend Kapitel 5. Die Position und Größe eines Objektes wird durch eine rechteckige Region beschrieben.

8.3.2 Bewertung eines semantischen Merkmals

Da mehrere semantische Merkmale in unterschiedlicher Größe in einem Bild enthalten sind, ist eine Bewertung der einzelnen Merkmale erforderlich. Insbesondere die Größe eines Merkmals nach der Skalierung des Bildes bestimmt dessen Bedeutung für das adaptierte Video. Durch das Abschneiden von Bildrändern ist ein Merkmal möglicherweise nicht mehr beziehungsweise nur noch zum Teil im Bild enthalten. Falls keine Ränder abgeschnitten werden sollen, erscheint das gesamte Bild in verkleinerter Darstellung, so dass Merkmale aufgrund ihrer geringen Größe nicht mehr erkannt werden könnten.

Anhand der analysierten Videosequenzen wird deutlich, dass eine Kombination aus Skalierung und dem Entfernen von Bildrändern im Allgemeinen zu den besten Ergebnissen führt. Abbildung 8.5 zeigt das Bild eines historischen Videos, in dem drei Bildregionen mit semantischen Inhalten automatisch erkannt werden. In dem Beispiel entstehen durch eine Skalierung oder das Abschneiden der Bildränder, wie auch in der Abbildung 8.5 (a) und (b) deutlich wird, adaptierte Videos in sehr schlechter Qualität. Es fehlen wichtige Teile des Bildes, oder die Bildinhalte können wegen der geringen Größe nicht mehr erkannt werden. Die Kombination beider Verfahren, also die Auswahl einer geeigneten Bildregion mit anschließender Skalierung, kann, wie es auch in den Abbildungen 8.5 (c) und (d) deutlich wird, zu einem wesentlich besseren Bild führen.

Das im Folgenden vorgestellte neue Verfahren bewertet die automatisch erkannten semantischen Merkmale, um einen möglichst guten Kompromiss zwischen einer Skalierung und dem Abschneiden der Ränder zu erreichen. Jedes semantische Merkmal wird durch eine rechteckige Region beschrieben. Es wird die Annahme getroffen, dass ein proportionaler Zusammenhang zwischen der Größe eines Merkmals und der Menge der dargestellten Informationen besteht. Dabei hängt die Bedeutung der Information von der Größe des Merkmals im adaptierten Video ab.

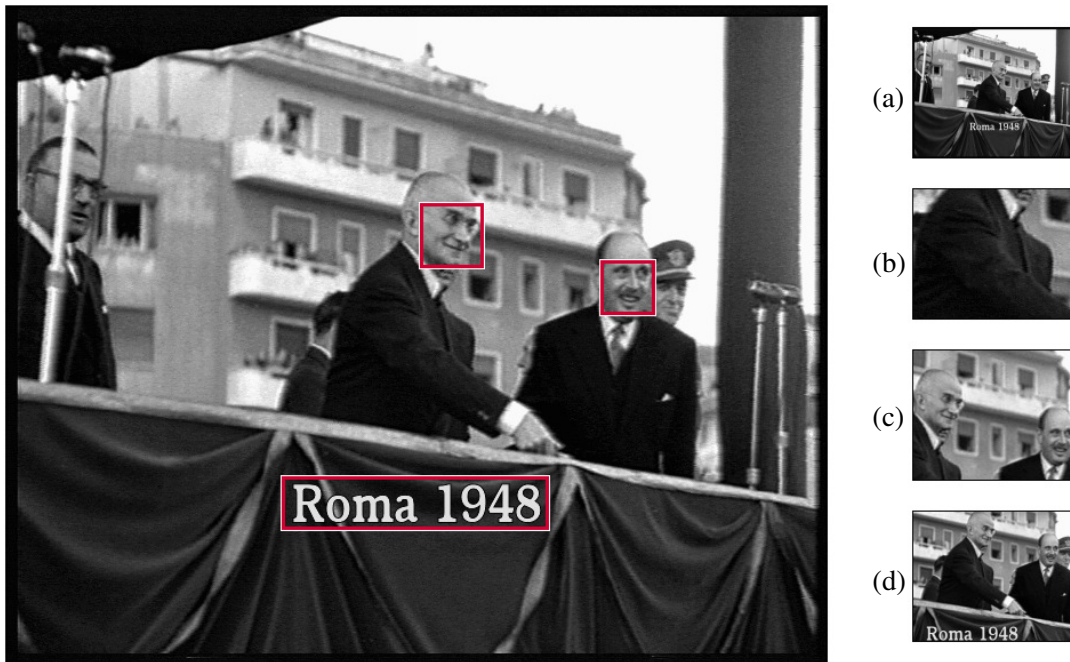


Abbildung 8.5: *Adaption der Bildauflösung durch Skalierung (a) und das Abschneiden von Rändern (b). Die Qualität der adaptierten Bilder steigt deutlich, falls zwei (c) oder drei (d) semantische Merkmale berücksichtigt werden.*

Für jedes semantische Merkmal ist eine *minimal zulässige Größe* (engl. *minimal perceptible size*) definiert. Falls die Größe des Merkmals durch die Skalierung unterschritten wird, kann der Inhalt nicht mehr oder nur noch eingeschränkt erkannt werden, und das Merkmal bleibt unberücksichtigt. Andererseits existiert eine Obergrenze für die Größe eines semantischen Merkmals, ab der kein zusätzlicher Nutzen für den Betrachter entsteht. Wird beispielsweise ein Text in einer akzeptablen Größe angezeigt, dann liefert eine größere Darstellung keine wichtigen zusätzlichen Informationen. Daher wird neben der *minimalen zulässigen Größe* auch eine *maximale sinnvolle Größe* für semantische Merkmale definiert.

Zur Bewertung der dargestellten Informationen wird eine Bildregion ausgewählt und auf die gewünschte Größe skaliert. Die Bewertung erfolgt anhand der identifizierten semantischen Merkmale innerhalb des skalierten Bildes, wobei nur die Merkmale berücksichtigt werden, die vollständig in der ausgewählten Bildregion liegen. Ist beispielsweise nur ein Teil eines Gesichtes oder einer Textzeile zu erkennen, so bleibt dieses Merkmal unberücksichtigt. Die Größe einer ausgewählten Bildregion darf die Auflösung des adaptierten Videos nicht unterschreiten, da eine Bildvergrößerung zu Unschärfe führen würde.

Der Wert zur Beschreibung des *Informationsgehaltes eines Merkmals* $V_i \in [0; 1]$ wird durch

die Größe des Merkmals i definiert:

$$V_i = \begin{cases} \frac{H_{max}}{H_i} & H_i > H_{max}, \\ \frac{H_i}{H_{max}} & H_{min} \leq H_i \leq H_{max}, \\ 0 & H_i < H_{min}. \end{cases} \quad (8.5)$$

Die Schwellwerte H_{min} und H_{max} legen die minimal zulässige bzw. maximal sinnvolle Größe eines Merkmals fest. Bei der manuellen Festlegung der beiden Schwellwerte sollten die Art des Displays, die Entfernung des Anwenders und individuelle Nutzerpräferenzen berücksichtigt werden.

Die Größe eines Merkmals wird durch die Höhe der rechteckigen Merkmalsregion beschrieben. Für Textregionen werden die Werte für H_{min} und H_{max} aus der Höhe des Zeichensatzes abgeleitet, für die der Text im adaptierten Video gut gelesen werden kann. Bei den anderen semantischen Merkmalen ist ein Schwellwert für die maximal sinnvolle Größe nicht erforderlich. Zur Berechnung der dargestellten Informationen für Gesichter oder Objekte wird H_{max} als Bildhöhe des ursprünglichen Videos festgelegt.

8.3.3 Auswahl und Kombination von Bildregionen

Die Größe und Position einer Bildregion wird so festgelegt, dass die Information innerhalb dieser Region maximal wird. Die gesamten Informationen $V_{sum}(R)$ werden durch die semantischen Merkmale innerhalb der ausgewählten Bildregion R bestimmt:

$$V_{sum}(R) = \sum_i S_i(R) \cdot V_i(R) \quad \text{mit} \quad (8.6)$$

$$V_i(R) = \begin{cases} \frac{H_{max}}{H_i(R)} & H_i(R) > H_{max}, \\ \frac{H_i(R)}{H_{max}} & H_{min} \leq H_i(R) \leq H_{max}, \\ 0 & H_i(R) < H_{min} \quad \text{und} \end{cases} \quad (8.7)$$

$$S_i(R) = \begin{cases} 1 & \text{falls } V_i \text{ vollständig in } R \text{ enthalten ist,} \\ 0 & \text{sonst.} \end{cases} \quad (8.8)$$

$V_i(R)$ bewertet die Information des semantischen Merkmals i in Abhängigkeit von der ausgewählten Bildregion R . $S_i(R)$ beschreibt in Form einer Binärvariablen, ob der Bildbereich des

semantischen Merkmals vollständig in der Region enthalten ist.

Die Überprüfung aller Positionen und Größen für die Bildregion R ist wegen der großen Anzahl an Kombinationen nicht sinnvoll. Sofern die maximal sinnvolle Größe unberücksichtigt bleibt, kann die Anzahl der zu analysierenden Regionen deutlich eingeschränkt werden. Damit $V_{sum}(R)$ maximal wird, muss jeder Rand der Bildregion R mit mindestens einem Rand eines semantischen Merkmals i übereinstimmen, und die Merkmale, die den Rand der Bildregion definieren, müssen vollständig in der Region enthalten sein. Es wird angenommen, dass für eine Bildregion beide Bedingungen erfüllt sind. Bei einer minimalen Verkleinerung der Bildregion würde mindestens ein semantisches Merkmal i nicht mehr vollständig in R enthalten, so dass der Wert von $V_{sum}(R)$ um $V_i(R)$ sinkt. Eine geringfügige Vergrößerung der Bildregion würde zu einer stärkeren Skalierung des Bildes führen, so dass die Werte aller Merkmale innerhalb der Bildregion sinken.

Falls nur Gesichter oder Objekte im Bild vorhanden sind, liefert das dargestellte Verfahren den optimalen Wert für $V_{sum}(R)$, da H_{max} der Bildhöhe entspricht. Bei Textregionen ist es wegen der maximal sinnvollen Größe möglich, dass ein kleinerer Text zu einem besseren Ergebnis für $V_{sum}(R)$ führt. Um dennoch eine Region mit einem möglichst hohen Informationsgehalt effizient zu ermitteln, wird zunächst die optimale Bildregion bestimmt, ohne die maximal sinnvolle Größe für Textregionen zu berücksichtigen. Anschließend wird die Bildregion bis zum maximalen Wert von $V_{sum}(R)$ vergrößert, wobei H_{max} in die Berechnung einfließt.

Ein effizientes Verfahren zur Berechnung der Bildregion wird im Folgenden vorgestellt: Zunächst werden einzelne Merkmale als Bildregion ausgewählt, und die Information dieser Region wird in Abhängigkeit von der erforderlichen Skalierung berechnet und gespeichert. Anschließend werden jeweils zwei Merkmale kombiniert, welche die Ränder der Bildregion festlegen. Das Verfahren wird fortgesetzt, bis für alle Kombinationen der Merkmale die Werte für $V_{sum}(R)$ bekannt sind. Die Region R mit dem maximalen Wert für $V_{sum}(R)$ definiert den Bildausschnitt des adaptierten Videos.

Durch die Kombination der Merkmale liegt die Komplexität des Algorithmus bei 2^N , wobei N die Anzahl der semantischen Merkmale eines Bildes angibt. Unter der Annahme, dass die Bildauflösung in Videos auf die Fernsehauflösung beschränkt ist, werden innerhalb der Bildfläche im Allgemeinen nur wenige semantische Merkmale erkannt. In den analysierten Videos liegt die tatsächliche maximale Anzahl bei $N = 5$, so dass im ungünstigsten Fall 32 Kombinationen überprüft werden. Um auch bei einer größeren Anzahl an semantischen Merkmalen eine schnelle Berechnung zu gewährleisten, werden die kleinsten Merkmalsregionen bei mehr als acht Merkmalen verworfen.

Das Verhältnis von Bildbreite zur Bildhöhe der ausgewählten Bildregion R entspricht im Allgemeinen nicht dem Verhältnis im adaptierten Video, so dass die Breite oder Höhe der ausgewählten Bildregion entsprechend vergrößert wird. In allen Bildern, in denen keine semantischen Merkmale identifiziert werden, wird das gesamte Bild als Bildregion verwendet, wobei schwarze Balken am Bildrand abgeschnitten werden.

8.3.4 Festlegung der Regionen für Kameraeinstellungen

Obwohl jede ausgewählte Region die dargestellten Informationen eines einzelnen Bildes maximiert, ist die Auswahl für Videos nicht gut geeignet, da plötzliche Größenänderungen und Sprünge innerhalb einer Kameraeinstellung auftreten, die als sehr störend empfunden werden. Schon kleinere Veränderungen der Position eines einzelnen semantischen Merkmals führen zu deutlich verwackelten Kameraeinstellungen.

Die Änderung der Position oder Größe der ausgewählten Bildregion soll innerhalb einer Kameraeinstellung kontinuierlich über mehrere Bilder erfolgen. Zunächst wird die Bildgröße angepasst, indem die Bildhöhe aller Bilder der Kameraeinstellung durch eine lineare Funktion angenähert wird. Die Bildbreite wird passend zum adaptierten Video festgelegt. Anschließend werden die Bildpositionen der ausgewählten Regionen geglättet, wobei die horizontalen und vertikalen Bildpositionen unabhängig voneinander durch eine lineare Funktion beschrieben werden. Bewegt sich beispielsweise in einer Aufnahme mit einer statischen Kamera ein einzelnes Objekt horizontal durch das Bild, so wird ein passender Schwenk erzeugt, durch den das Objekt während der gesamten Kameraeinstellung im Bildzentrum liegt.

Drei Fälle werden besonders berücksichtigt, bei denen die Glättung der Größen- und Positionswerte nicht zu zufriedenstellenden Ergebnissen führt. Falls in einem einzelnen Bild ein Merkmal falsch oder gar nicht erkannt wird, entstehen deutliche Fehler bei der linearen Annäherung der Positionen und Größen der ausgewählten Regionen. Daher werden die Bilder, in denen die Größe oder Position einer Region deutlich von den Regionen der benachbarten Bilder abweicht, bei der Berechnung nicht berücksichtigt.

Der zweite Fall tritt insbesondere bei längeren Kameraeinstellungen auf, in denen Objekte im Bild erscheinen oder verschwinden. Kameraeinstellungen mit einer Länge von mehr als 30 Sekunden werden in zwei Abschnitte unterteilt, wobei die Grenzen so festgelegt werden, dass die ausgewählten Bildregionen innerhalb der Abschnitte möglichst ähnlich sind. Die Berechnung der linearen Funktionen erfolgt für die einzelnen Abschnitte unabhängig voneinander. Um eine plötzliche Änderung der Bewegung der Kamera beim Übergang zweier Abschnitte

zu vermeiden, werden die Positions- und Größenwerte im Bereich der Übergänge durch einen Gaußfilter geglättet.

Es ist möglich, dass zwei räumlich getrennte Bildregionen eine sehr ähnliche Bewertung erhalten. In diesem Fall wird nur die Region mit der maximalen Information ausgewählt, die andere Region bleibt unberücksichtigt. Um die dargestellten Informationen innerhalb einer Kameraeinstellung zu erhöhen, wird ein ähnlicher Ansatz wie beim *Photo2Video*-System vorgeschlagen [216]. Ziel des Systems ist es, aus einem Foto ein Video zu erzeugen, in dem wichtige Bildinhalte nacheinander im Detail dargestellt werden. Zur Erzeugung des Videos können komplexe Kamerabewegungen wie beispielsweise ein Schwenk kombiniert mit einem Zoomeffekt verwendet werden.

Der direkte Ansatz von *Photo2Video* wurde zur Umwandlung von Fotos entwickelt und ist für die Adaption von Videos nur bedingt geeignet. Für die Adaption von Videos ist es wichtig, dass die Dauer einer Kameraeinstellung unverändert bleibt. Zudem sind bei der direkten Umsetzung des *Photo2Video*-Ansatzes in mehreren Testvideos komplexe Kamerabewegungen in aufeinander folgenden Kameraeinstellungen aufgetreten, die beim Betrachten als unangenehm empfunden werden.

Der neue im Folgenden vorgestellte Ansatz berücksichtigt diesen Sachverhalt und erzeugt künstliche Kamerabewegungen, ohne die Länge des Videos zu verändern. Es wird angenommen, dass zwei relevante Bildregionen in einer Kameraeinstellung erkannt wurden. Damit möglichst viele Bildinhalte im adaptierten Video erhalten bleiben, wird zufällig eine der beiden Regionen als erstes Bild der Kameraeinstellung festgelegt, die andere Region definiert den Bildausschnitt für das letzte Bild der Kameraeinstellung. Bei räumlich benachbarten Bildregionen wird ein linearer Übergang zwischen den Regionen berechnet, so dass ein künstlicher Kameraschwenk entsteht. Ansonsten wird die Kameraeinstellung durch einen harten Schnitt unterteilt. Eine künstliche Zoomoperation wird eingefügt, falls eine kleine Bildregion in einer sehr langen Kameraeinstellung (> 30 Sekunden) ausgewählt wird. Die Bildregion im ersten oder letzten Bild der Kameraeinstellung wird auf die Bildgröße des Videos gesetzt und ein linearer Übergang zwischen den Bildregionen des ersten und des letzten Bildes erzeugt, so dass eine Zoomoperation innerhalb der Kameraeinstellung entsteht.

Nachdem die Bildregionen für alle Kameraeinstellungen des Videos spezifiziert sind, werden sie auf die gewünschte Größe durch lineare Interpolation mittels Gleichung 4.1 skaliert und zusammen mit der Audiospur als Video kodiert und gespeichert.

8.4 Anpassung der Bildqualität historischer Videos

Bei dem in Kapitel 2.3.6 vorgestellten Projekt *European Chronicles Online* wurde eine komplexe Anwendung zur Verwaltung und Indexierung von historischen Videoarchiven entwickelt. Die Bildqualität der in diesem Archiv gespeicherten historischen Schwarz-Weiß-Filme ist mit der Qualität aktueller Filme nicht vergleichbar. Durch die Lagerung der Filmrollen über mehrere Jahrzehnte und den mechanischen Abrieb bei der Projektion der Filme sind viele Bildfehler in den Videos entstanden. Beim Betrachten eines Videos wird eine gewisse Qualität erwartet, die insbesondere bei stark verwackelten oder schlecht belichteten Aufnahmen nicht gegeben ist. Algorithmen zur Adaption der Bildqualität ermöglichen es, die Darstellung eines historischen Videos zu verbessern.

Ziel der Anpassung der Bildqualität von historischen Videos ist es, typische Bildfehler zu identifizieren und zu korrigieren. *Helligkeitsschwankungen* innerhalb eines kürzeren Zeitraums sowie über- oder unterbelichtete Kameraeinstellungen werden durch eine Anpassung der durchschnittlichen Helligkeit und eine Erhöhung des Kontrastes ausgeglichen. Fehler in Form von *hellen Streifen* entstehen durch den Abrieb beim Filmtransport mit den alten Projektoren. Zur Korrektur werden die fehlerhaften Pixel durch benachbarte Pixelwerte interpoliert. Stark *verwackelte Kameraeinstellungen* fallen beim Betrachten eines Videos negativ auf und werden anhand der Kamerabewegung identifiziert und ausgeglichen. In der Abbildung 8.8 sind Beispiele für Videosequenzen mit Bildfehlern abgebildet, die von Algorithmen zur automatischen Verbesserung der Bildqualität deutlich profitieren.

8.4.1 Korrektur der Helligkeit in historischen Videos

In historischen Videos sind deutliche *Helligkeitsschwankungen* innerhalb kurzer Zeiträume möglich, die zu einem Flackern des Bildes führen. Die Helligkeitsänderungen entstehen durch die mangelhafte Technik der Projektoren und die Lagerung der Filme über mehrere Jahrzehnte [455]. Um Helligkeitsschwankungen zu erkennen, wird die durchschnittliche Helligkeit I_i der Pixel eines Bildes i berechnet. Falls innerhalb einer Kameraeinstellung das Maximum I_{max} der durchschnittlichen Helligkeit eines Bildes deutlich über dem Minimum I_{min} liegt, soll die Helligkeit korrigiert werden.

Zunächst wird die Helligkeit der Bilder an die durchschnittliche Helligkeit I_{avg} der Kameraeinstellung angeglichen. Der Korrekturfaktor $F_I(i)$ definiert die absolute Helligkeitsänderung aller Pixel eines Bildes i :

$$F_I(i) = \alpha \cdot (I_{avg} - I_i). \quad (8.9)$$

Ein Skalierungsfaktor von $\alpha = 1$ führt zu einer vollständigen Korrektur der Helligkeit. Da innerhalb einer Kameraeinstellung der Kontrast in besonders dunklen oder hellen Bildern niedriger ist als in den übrigen Bildern, erscheinen diese Bilder nach der Anpassung der Helligkeit sehr kontrastarm. Ein Skalierungsfaktor von $\alpha = 0,8$ führt in Kombination mit einer Erhöhung des Kontrastes $F_C(i)$ zu deutlich besseren Ergebnissen:

$$F_C(i) = \beta \cdot |F_I(i)|. \quad (8.10)$$

Durch die Anpassung des Kontrastes mit einem Skalierungsfaktor von $\beta = 0,5$ wird die noch verbleibende Helligkeitsdifferenz ausgeglichen und ein kontrastreiches Bild erzeugt.

Neben den Helligkeitsschwankungen sind in den historischen Videos auch stark *über-* oder *unterbelichtete Kameraeinstellungen* enthalten, in denen die Bildinhalte nur sehr schwer erkannt werden können. Diese Kameraeinstellungen entstehen durch Fehler bei der Aufnahme oder eine falsche Entwicklung der Filme. Die Anpassung der durchschnittlichen Helligkeit wird am Beispiel zu dunkler Kameraeinstellungen erläutert. Falls die durchschnittliche Helligkeit des hellsten Bildes I_{max} unter einem Schwellwert liegt, gilt die Kameraeinstellung als zu dunkel, und eine Anpassung der Helligkeit ist erforderlich. Im Rahmen der Analyse der adaptierten historischen Videos wurde deutlich [285], dass zur Korrektur der durchschnittlichen Helligkeit lediglich eine Erhöhung des Kontrastes entsprechend der Gleichung 8.10 erforderlich ist.

8.4.2 Korrektur von Streifen und Kratzern im Bild

Horizontale oder vertikale Streifen entstehen bei der Entwicklung eines Filmes oder durch den mechanischen Abrieb beim Transport der Filmrolle. Die überwiegend hellen Streifen sind unabhängig vom Bildinhalt über einen längeren Zeitraum sichtbar, so dass die Erkennung und Korrektur der Streifen nicht auf einzelne Kameraeinstellungen beschränkt wird. Bezogen auf die Bildhöhe bzw. Bildbreite variiert die Position eines Streifen in den analysierten historischen Videos um maximal fünf Prozent [455].

In einem zweistufigen Analyseprozess werden zunächst alle horizontalen und vertikalen Linien im Bild identifiziert, unter denen auch echte Bildinhalte enthalten sein können, die nicht

korrigiert werden sollen. Die Erkennung eines Kratzers erfolgt durch eine Analyse der Linienpositionen im Zeitablauf.

Die Erkennung wird beispielhaft für horizontale Streifen erläutert. Dazu wird das Bild zeilenweise durchlaufen und für jede Zeile die Anzahl der Linienpixel gespeichert. Ein Pixel zählt als *Linienpixel*, falls die Helligkeit einen Schwellwert übersteigt, in der horizontalen Nachbarschaft weitere helle Pixel liegen und eine helle Fläche durch einen Vergleich mit den vertikal benachbarten Pixeln ausgeschlossen werden kann. Übersteigt die Anzahl der Linienpixel in einer Zeile einen Schwellwert, so wird die Position dieser Zeile als möglicher Kratzer gespeichert. Im einem zweiten Schritt werden alle Streifen verworfen, bei denen in benachbarten Bildern an ähnlichen Positionen nur selten Streifen vorkommen.

Die durch Kratzer verursachten Bildfehler verändern eine einzelne Zeile oder Spalte deutlich und beeinflussen die angrenzenden Pixel nur geringfügig. Zur Korrektur eines horizontalen Streifens wird der Wert eines Linienpixels $I_{x,y}$ mit dem Durchschnittswert der vertikal indirekt benachbarten Pixel ersetzt:

$$I'_{x,y} = \frac{1}{2} \cdot (I_{x,y-2} + I_{x,y+2}). \quad (8.11)$$

Da in den direkt angrenzenden Zeilen $(y-1)$ und $(y+1)$ Fehler enthalten sein können, werden auch die Pixel dieser Zeile durch eine Gewichtung der ursprünglichen Helligkeitswerte mit dem Faktor $\gamma \in [0, 1]$ angepasst:

$$I'_{x,y-1} = \gamma \cdot I_{x,y-1} + (1 - \gamma) \cdot I_{x,y-2} \quad \text{und} \quad (8.12)$$

$$I'_{x,y+1} = \gamma \cdot I_{x,y+1} + (1 - \gamma) \cdot I_{x,y+2}. \quad (8.13)$$

In experimentellen Ergebnissen hat sich ein Gewichtungsfaktor von $\gamma = 0,25$ als geeigneter Wert herausgestellt [285].

8.4.3 Korrektur verwackelter Kameraeinstellungen

Ein weiterer typischer Fehler in historischen Videos sind stark *verwackelte Sequenzen*. Eine Kameraeinstellung gilt als verwackelt, falls sich die Bilder innerhalb eines kurzen Zeitraums horizontal oder vertikal zuerst in die eine und dann in die andere Richtung bewegen. Diese Fehler sind auf einen ungleichmäßigen Filmtransport bei der Aufnahme zurückzuführen. Die

in diesem Abschnitt vorgestellten Algorithmen eignen sich auch zur Qualitätsverbesserung von Amateurvideos, die ohne Stativ aufgenommen wurden und stark verwackelt sind.

Um verwackelte Kameraeinstellungen zu erkennen, wird die Kamerabewegung zwischen benachbarten Bildern mit Hilfe des in Kapitel 3 vorgestellten Verfahrens berechnet. Durch Analyse der Parameter t_x und t_y der Gleichung 3.1, welche die horizontale und vertikale Verschiebung des Bildes beschreiben, erfolgt die Erkennung von verwackelten Kameraeinstellungen. Innerhalb einer Kameraeinstellung darf die Summe der horizontalen und vertikalen Kamerabewegungen M_H bzw. M_V nur geringfügig von null abweichen:

$$M_H = \frac{1}{N_L - N_F} \cdot \sum_{j=N_F}^{N_L-1} t_x(j) \quad \text{und} \quad (8.14)$$

$$M_V = \frac{1}{N_L - N_F} \cdot \sum_{j=N_F}^{N_L-1} t_y(j). \quad (8.15)$$

N_F und N_L spezifizieren das erste bzw. letzte Bild einer Kameraeinstellung. $t_x(j)$ beschreibt die horizontale Translation zwischen Bild j und $j + 1$. Neben der Bedingung, dass bei einer verwackelten Kameraeinstellung M_H und M_V nahe bei null liegen, muss die Summe der durchschnittlichen absoluten horizontalen oder vertikalen Bewegungen M_H^{abs} bzw. M_V^{abs} deutlich über null liegen:

$$M_H^{abs} = \frac{1}{N_L - N_F} \cdot \sum_{j=N_F}^{N_L-1} |t_x(j)| \quad \text{und} \quad (8.16)$$

$$M_V^{abs} = \frac{1}{N_L - N_F} \cdot \sum_{j=N_F}^{N_L-1} |t_y(j)|. \quad (8.17)$$

Zur *Korrektur einer verwackelten Kameraeinstellung* wird das mittlere Bild einer Kameraeinstellung als Referenzbild festgelegt und die übrigen Bilder entsprechend ausgerichtet. Die Verschiebung wird auf ganzzahlige Werte für t_x und t_y eingeschränkt, da sonst eine lineare Interpolation der Pixelwerte erforderlich wird und das Bild unscharf wird. Durch die Verschiebung entstehen Bereiche an den Bildrändern ohne verfügbare Bildinformationen. Diese Randbereiche werden durch schwarze Pixel ersetzt und für alle Bilder der Kameraeinstellung übernommen. Alternativ besteht die Möglichkeit, die Randbereiche durch Bildinformationen aus vorhergehenden oder folgenden Bildern zu ersetzen, wobei durch Objektbewegungen auf-

fällige Verzerrungen in diesen Bereichen entstehen können.

8.5 Experimentelle Ergebnisse

Im Rahmen der experimentellen Ergebnisse werden die unterschiedlichen Adaptionsverfahren analysiert. In Abbildung 8.6 werden Ergebnisse zur *Adaption der Farbtiefe* am Beispiel von Binärbildern für zwei Videosequenzen vorgestellt. Die Umwandlung in ein Binärbild erfolgt in Abbildung 8.6 (Mitte) durch einen Vergleich mit einem festen Schwellwert. Obwohl dieser Schwellwert manuell und somit optimal festgelegt wurde, können nur wenige Objekte gut erkannt werden, und große Bildbereiche enthalten keine Informationen. Um Videos mit diesem Verfahren automatisch zu adaptieren, müsste zusätzlich ein geeigneter Schwellwert geschätzt werden, so dass eine Verschlechterung der Bildqualität zu erwarten ist.

Zum Vergleich sind in der Abbildung 8.6 (rechts) die entsprechenden Bilder des neuen Adaptionsverfahrens gegenübergestellt. Durch die Überlagerung mit texturierten Binärbildern können deutlich mehr Bildinhalte sowohl bei einzelnen Vordergrundobjekten als auch im Bildhintergrund erkannt werden. Da bei der Adaption variable Intervallgrößen verwendet werden und eine manuelle Festlegung von Schwellwerten nicht erforderlich ist, sind gute Ergebnisse sowohl für sehr helle und dunkle als auch für sehr kontrastarme Kameraeinstellungen möglich [284].

Ergebnisse zur *Adaption der Bildauflösung* [280] werden am Beispiel von zwei historischen Videos verdeutlicht. Bilder ausgewählter Kameraeinstellungen des ursprünglichen, lediglich auf die passende Bildgröße skalierten Videos sind in der Abbildung 8.7 (a) dargestellt. In den skalierten Bildern können wichtige Bildinhalte wegen ihrer geringen Größe teilweise nicht mehr erkannt werden. Für die adaptierten Bilder in der Abbildung 8.7 (b) erfolgt die Auswahl der Bildregionen anhand semantischer Merkmale.

In der ersten Videosequenz aus dem Jahre 1947 ist im unteren Bildbereich ein Zeitcode eingeblendet. Obwohl dieser als Textregion identifiziert wird, bleibt er unberücksichtigt, da die minimal zulässige Größe unterschritten wird. In der ersten Kameraeinstellung definieren die drei Textzeilen die Position und Größe der ausgewählten Bildregion. In der zweiten überdurchschnittlich langen Kameraeinstellung wird ein künstlicher Zoom auf das Gesicht der Person erzeugt, so dass im letzten Bild dieser Kameraeinstellung das Gesicht in voller Größe zu sehen ist. Der Algorithmus zum Auffinden von Gesichtsregionen erkennt das Gesicht in der dritten Kameraeinstellung vermutlich aufgrund der Brille und des Bartes nicht. Daher wird das vollständige Bild angezeigt, wobei ein kleiner fehlerhafter Randbereich mit schwarzen bzw.



Abbildung 8.6: Beispiele zweier Videos zur Adaption der Farbtiefe: Originalbild (links), Binärbild nach Vergleich mit einem optimalen (manuell festgelegten) Schwellwert (mitte) und automatisch erzeugtes Binärbild durch Überlagerung von Texturen und einer Verstärkung der Kanten (rechts).

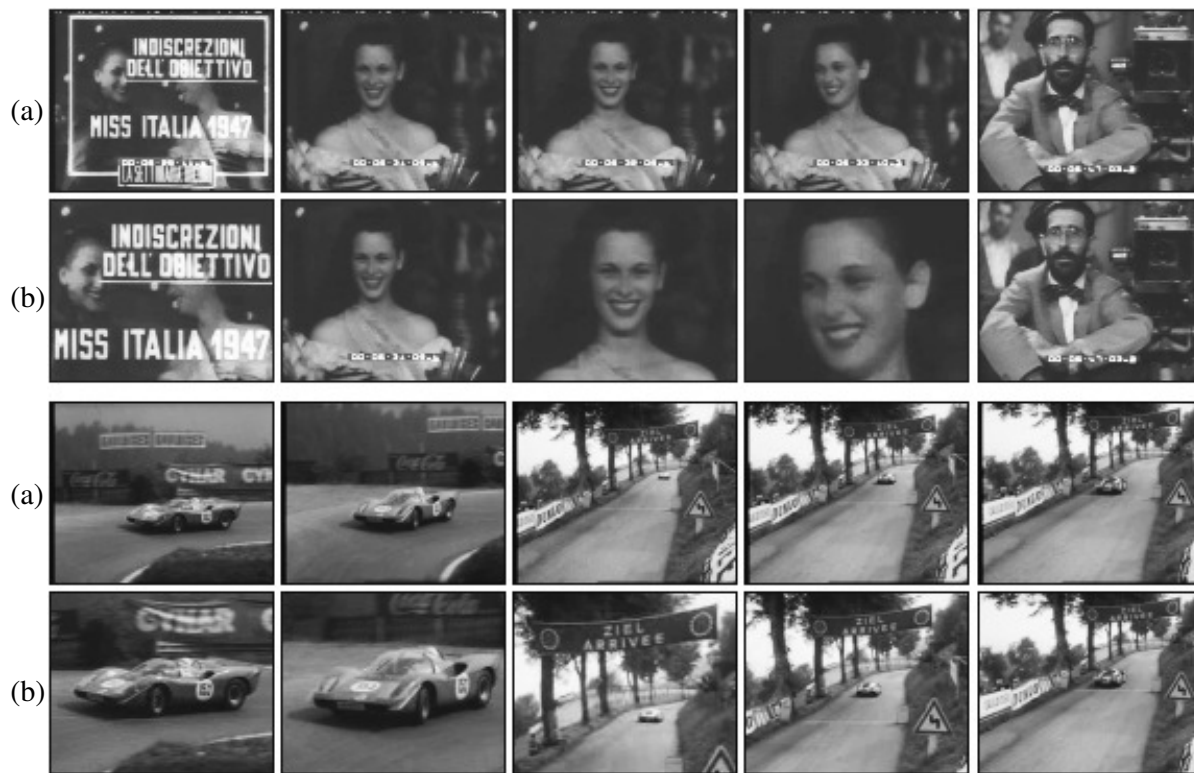


Abbildung 8.7: Beispiele zur Adaption der Bildauflösung für zwei historische Filme: Skalierte Originalvideos (a) und Videos nach semantischer Adaption der Bildauflösung (b)

verrauschten Pixeln abgeschnitten wird.

Beim zweiten Video handelt es sich um ein historisches Autorennen, in dem Rennwagen identifiziert und im adaptierten Video hervorgehoben werden. In der ersten Kameraeinstellung wird die Bildregion anhand der Position des Fahrzeugs bestimmt und die Größe des Bildausschnittes durch die Breite des Rennwagens definiert. Das Fahrzeug erscheint im semantisch adaptierten Video leicht nach rechts versetzt, da auch der Schatten des Rennwagens segmentiert wird. In der zweiten Kameraeinstellung wird ein künstlicher ausgehender Zoom eingefügt, so dass in den ersten Bildern Details wie beispielsweise der Text über der Ziellinie noch erkannt werden können.

Die Anpassung fehlerhafter historischer Videos wird am Beispiel von 20 kurzen Videosequenzen mit einer Länge zwischen 30 und 120 Sekunden analysiert. Neun Sequenzen enthalten deutliche Helligkeitsschwankungen, die durch die Anpassung der durchschnittlichen Helligkeit und die Erhöhung des Kontrastes so gut ausgeglichen werden, dass sie nach der Adaption nicht mehr wahrgenommen werden. In Abbildung 8.8 wird die Adaption historischer Videos

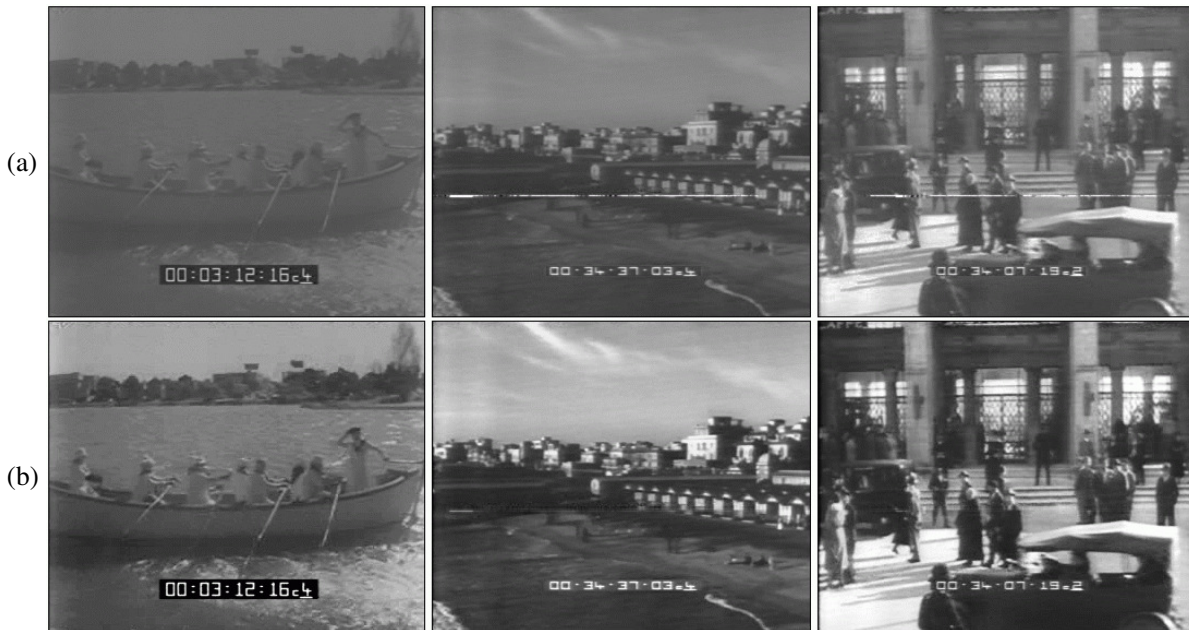


Abbildung 8.8: *Adaption historischer Videos durch Anpassung der Helligkeit und Korrektur von Kratzern: Originalvideo (a) und adaptiertes Video (b)*

an drei Beispielen verdeutlicht. Insbesondere in den ersten beiden Kameraeinstellungen der Abbildung 8.8 sind die Bildinhalte nach der Adaption deutlich besser zu erkennen.

Sechs Sequenzen enthalten horizontale Streifen, die in mehr als 95 Prozent der Bilder korrekt identifiziert werden. Fehler treten insbesondere bei mehrfach unterbrochenen oder sehr kurzen Kratzern auf. Durch die Überprüfung der Position eines Streifens im Zeitablauf wird sichergestellt, dass keine echten Bildinhalte als Linie erkannt werden. Beispiele für die Korrektur eines Streifens sind im zweiten und dritten Bild der Abbildung 8.8 zu sehen.

Durch die hohe Zuverlässigkeit bei der Berechnung der Kameraparameter ist eine nahezu fehlerfreie Identifikation und Korrektur der verwackelten Kameraeinstellungen möglich. Der Aufwand für die Korrektur der Kamerabewegung ist im Vergleich zu den anderen Verfahren sehr hoch. Zur Verringerung des Rechenaufwandes werden zunächst verwackelte Kameraeinstellungen identifiziert, indem die Kamerabewegung zwischen zehn aufeinander folgenden Bildern berechnet wird. Nur im Falle einer verwackelten Kameraeinstellung erfolgt die Berechnung für die übrigen Bilder. In allen acht Sequenzen, die verwackelte Kameraeinstellungen enthalten, wird die Bildqualität deutlich verbessert, und das Bild erscheint wesentlich stabiler.

8.6 Zusammenfassung

In diesem Kapitel wurden neue Verfahren zur semantischen Adaption von Videos vorgestellt. Die Adaption der Farbtiefe erfolgte durch Analyse der Helligkeitsverteilung innerhalb der gesamten Kameraeinstellung. Zudem wurde ein neues Adaptionsverfahren zur Erzeugung von Binärbildern entwickelt, das Kanteninformationen mit Texturen kombiniert und auch für die Adaption von Videos geeignet ist. Ein weiteres neues Verfahren zur Adaption der Bildauflösung wurde vorgestellt, bei dem semantische Inhalte des Videos analysiert, bewertet und zu Regionen zusammengefasst werden. Ein Algorithmus zur geeigneten Auswahl von Regionen in Kameraeinstellungen wurde entwickelt, durch den künstliche Kamerabewegungen und Kameraoperationen eingefügt werden, um die Bildinhalte im adaptierten Video besser darzustellen. Ein drittes neues Adaptionsverfahren wurde zur Verbesserung der Bildqualität von historischen Videos entwickelt und ermöglicht die automatische Korrektur der Helligkeit und des Kontrastes, die Entfernung von Streifen und Kratzern sowie die Stabilisierung verwackelter Aufnahmen.

Für viele Videos bietet sich eine Kombination der drei in diesem Kapitel vorgestellten Adaptionsverfahren an. Häufig ist bei einer Verringerung der Farbtiefe auch eine Anpassung der Bildauflösung erforderlich. Die Verfahren zur Adaption historischer Videos eignen sich auch für Amateurvideos, da in diesem Umfeld zunehmend Probleme mit der Bildqualität durch die Lagerung der Bänder auftreten. Automatische Verfahren zur Adaption von Videos sind auch für Filmarchive besonders interessant, die Videos einer breiten Öffentlichkeit über das Internet zur Verfügung stellen wollen. Eine Kombination der automatischen Adaptionsverfahren zur Verringerung der Bildauflösung und der Korrektur der Bildinhalte kann die Videos in geeigneter Form aufbereiten, ohne dass ein manuelles Bearbeiten der umfangreichen Filmsammlungen erforderlich ist.

KAPITEL 9

Computergenerierte Zusammenfassungen von Videos

Mit der Entwicklung immer leistungsfähigerer Computer ist neben Texten, Bildern und Audiodateien auch die Anzeige und Bearbeitung von digitalen Videos für Privatanwender möglich geworden. Gleichzeitig steigt der Umfang der verfügbaren digitalen Videos, da Fernsehanstalten sowie öffentliche und private Filmarchive ihre Filmsammlungen digitalisieren und über das Internet einer breiten Öffentlichkeit zur Verfügung stellen. Die Bedeutung von Videoarchiven, die eine Navigation und Suche in Videos unterstützen, nimmt mit dem Umfang der verfügbaren Videos kontinuierlich zu.

Im Vergleich zur Suche innerhalb eines Textdokumentes ist die Komplexität der Suche in Videos deutlich höher. Der Wechsel des Mediums von einem kontinuierlichen Medienstrom zu einer textuellen Beschreibung erfordert neue Suchstrategien für Videosequenzen. Anhand der Suchergebnisse kann ein Anwender nur sehr schwer erkennen, ob und welche der gefundenen Segmente des Videos seinen Erwartungen entsprechen. Nur durch das sehr zeitaufwendige Betrachten des Videos können die Inhalte im Detail aufgenommen werden.

Dieser hohe zeitliche Aufwand kann durch spezielle Methoden zur *schnellen Navigation* innerhalb eines Videos verringert werden, die im Folgenden beschrieben werden. Neben dem schnellen Abspielen in Vorwärts- und Rückwärtsrichtung ist ein direkter Sprung an eine beliebige Position innerhalb des digitalen Videos möglich. *Repräsentative Bilder* (engl. *key frame*) können dabei als Verweise auf Kameraeinstellungen innerhalb des Videos dienen. Obwohl diese Navigationstechniken die Zeit zum Auffinden spezieller Bereiche reduzieren, bleibt der

Zeitaufwand, um einen Überblick über das ganze Video zu erhalten, sehr hoch. Ein wichtiges Segment kann erst dann als solches identifiziert werden, wenn der entsprechende Abschnitt des Videos betrachtet wurde.

Durch eine intelligente automatische Auswahl und Kombination von Kameraeinstellungen kann eine *automatisch erzeugte Zusammenfassung eines Videos* (engl. *video summary*, *video abstract* oder *video skim*) einem Betrachter die wesentlichen Inhalte in kurzer Zeit vermitteln. Hierbei ist wichtig, dass der semantische Inhalt des Originalvideos in der deutlich kürzeren Zusammenfassung möglichst gut erhalten bleibt.

Um wichtige Segmente eines Videos von unwichtigen zu unterscheiden, werden Merkmale zur Beschreibung der einzelnen Kameraeinstellungen ermittelt. Besonders wichtige Kameraeinstellungen werden ausgewählt und zu einer Zusammenfassung kombiniert. Die *Darstellung der Zusammenfassung* kann *statisch* in Form einzelner repräsentativer Bilder oder *dynamisch* als Kombination von Kameraeinstellungen erfolgen.

Im Rahmen des in Kapitel 2.3.6 vorgestellten Projektes *European Chronicles Online* wurde ein komplexes Softwaresystem entwickelt, um große Archive mit historischen Videos zu verwalten und die historisch wertvollen Dokumentationen den Archivaren und der Öffentlichkeit leichter zugänglich zu machen. Eine besondere Herausforderung lag darin, die Inhalte der Videos sinnvoll darzustellen und eine effiziente Suche zu ermöglichen. Zur Unterstützung der Suche werden *Metadaten* – also zusätzliche Daten zur Beschreibung der Videos – im System gespeichert. Eine textbasierte Suchanfrage liefert als Ergebnis im *European-Chronicles-Online*-System eine Liste mit ausgewählten Einzelbildern, die durch textuelle Informationen ergänzt werden. Da der dynamische Charakter des Videos nicht berücksichtigt wird, gehen wichtige semantische Informationen bei dieser Form der Darstellung verloren. Kurze prägnante Zusammenfassungen in Form eines Videos können wesentlich dazu beitragen, den Inhalt des deutlich längeren Originalvideos schneller zu erfassen und die Arbeit mit umfangreichen Videoarchiven zu erleichtern. Die Algorithmen zur automatischen Erzeugung von Zusammenfassungen sind im Rahmen dieser Arbeit in das *European-Chronicles-Online*-System eingeflossen, so dass für jedes neu ins Archiv aufgenommene Video zusätzlich eine wesentlich kürzere Version als Zusammenfassung zur Verfügung gestellt wird.

In diesem Kapitel werden neue Verfahren zur automatischen Erzeugung von Zusammenfassungen vorgestellt, welche die besonderen Herausforderungen historischer Dokumentationen berücksichtigen. Um zu verhindern, dass bei der Auswahl repräsentativer Bilder für Kameraeinstellungen einzelne fehlerhafte Bilder ausgewählt werden, wird ein neuer Algorithmus vorgestellt, der die Ähnlichkeit des ausgewählten Bildes zu allen Bildern der Kameraeinstel-

lung berücksichtigt. Neben einzelnen fehlerhaften Bildern sind in den historischen Videos auch häufig fehlerhafte Kameraeinstellungen enthalten, die automatisch erkannt und ausgefiltert werden müssen. Dazu schlagen wir einen neuen Algorithmus vor, durch den Gruppen mit ähnlichen Kameraeinstellungen gebildet und gleichzeitig fehlerhafte Kameraeinstellungen identifiziert werden.

Weiterhin stellen wir mehrere neue Heuristiken zur Bewertung einzelner Merkmale vor: In die Bewertung der Kamerabewegung gehen die Art, Intensität und Dauer der Bewegung ein. Die Bewertung ähnlicher Kameraeinstellungen wird insbesondere durch die schon ausgewählten Kameraeinstellungen beeinflusst. Die Heuristik zur Auswahl von Kameraeinstellungen innerhalb einer Szene begünstigt die Wahl zweier benachbarter Kameraeinstellungen.

Nach der Bewertung von Merkmalen werden neue Algorithmen und Heuristiken zur Auswahl und Kombination relevanter Kameraeinstellungen präsentiert, welche die Eigenschaften von historischen Video-Dokumentationen berücksichtigen. Wesentliche Bestandteile umfassen die Erkennung nicht relevanter Kameraeinstellungen sowie den Einsatz von sowohl festen als auch dynamisch während des Auswahlprozesses veränderlichen Merkmalswerten. Als letzter Schritt erfolgt die Überprüfung der ausgewählten Kameraeinstellungen anhand spezieller Regeln. Am Beispiel einer Kollage wird eine neue Darstellungsform zur Präsentation statischer Zusammenfassungen erläutert. Abschließend wird auf Evaluationsergebnisse mit professionellen Nutzern von Videoarchiven eingegangen.

Im folgenden Abschnitt werden zunächst unterschiedliche Verfahren zur automatischen Erzeugung von Zusammenfassungen vorgestellt. Anschließend wird in Abschnitt 9.2 ein Überblick über das im Rahmen des *European-Chronicles-Online*-Projektes von uns entwickelte System zur Erzeugung von Zusammenfassungen für historische Video-Dokumentationen gegeben. Die Schwerpunkte der darauf folgenden Abschnitte liegen in der Berechnung geeigneter Merkmale zur Beschreibung von Kameraeinstellungen sowie in der Heuristik zur Auswahl der Kameraeinstellungen für die Zusammenfassung. Innerhalb der experimentellen Ergebnisse in Abschnitt 9.6 wird die Qualität der computergenerierten Zusammenfassungen für historische Videos aus dem *European-Chronicles-Online*-Projekt analysiert.

9.1 Heuristiken zur Erzeugung von Zusammenfassungen

Aus den Präferenzen des Betrachters und der Art des Filmmaterials wird abgeleitet, welche Informationen in einer Zusammenfassung kombiniert werden sollten. Dabei sind zwei Arten von Zusammenfassungen möglich. Bei einer *Vorschau eines Videos* (engl. *trailer*) soll die

Aufmerksamkeit und das Interesse eines Zuschauers gewonnen werden. Diese Zusammenfassung wird überwiegend für Spielfilme und Sportereignisse eingesetzt und fesselt die Zuschauer durch Kameraeinstellungen mit starken Emotionen, hoher Spannung und besonderen Ereignissen.

Die zweite Art der Zusammenfassung versucht, einen *Überblick über das Video* zu vermitteln. In kompakter Form werden die wesentlichen Inhalte aggregiert dargestellt, so dass diese Art der Zusammenfassung besonders gut für Dokumentationen und Nachrichtensendungen geeignet ist. Um einen guten Überblick zu geben, muss insbesondere die *Struktur des Videos* berücksichtigt werden, damit Wiederholungen und ähnliche Kameraeinstellungen nicht mehrfach in die Zusammenfassung einfließen [321, 414].

Der erste Schritt bei der Erzeugung einer Zusammenfassung beinhaltet die Einteilung in Segmente. Auf der visuellen Ebene eignen sich Kameraeinstellungen, innerhalb der Audiospur sind ruhige Bereiche zur Unterteilung des Videos besonders gut geeignet. In einem zweiten Schritt werden Szenen, Dialoge und Kameraeinstellungen mit visueller Ähnlichkeit identifiziert, um die Struktur des Videos abzuleiten. Eine Zusammenfassung sollte besonders relevante Kameraeinstellungen enthalten, wobei die Bedeutung der einzelnen Kameraeinstellungen aus den automatisch berechneten strukturellen und semantischen Informationen abgeleitet wird. Der letzte Schritt umfasst die Zusammenstellung, Speicherung und Präsentation der ausgewählten Inhalte. Dabei kann das Video als *statische* Zusammenfassung in Form von Einzelbildern oder *dynamisch* als Video mit deutlich reduzierter Dauer dargestellt werden. Abbildung 9.1 verdeutlicht die wesentlichen Schritte bei der Erzeugung einer Zusammenfassung.

Mehrere Publikationen wurden in den letzten Jahre veröffentlicht, in denen Verfahren zur automatischen Erzeugung von Zusammenfassungen für Videos vorgestellt werden [114, 310, 370, 396, 486]. Im Folgenden Abschnitt werden zunächst Merkmale aufgeführt, welche die Auswahl geeigneter Kameraeinstellungen für eine Zusammenfassung unterstützen. Anschließend werden bekannte Verfahren zur Erzeugung statischer und dynamischer Zusammenfassungen vorgestellt.

9.1.1 Allgemeine Merkmale zur Beschreibung von Kameraeinstellungen

Sowohl bei einer Vorschau eines Spielfilms als auch bei einem Überblick einer Dokumentation ist ein Zuschauer an den Höhepunkten interessiert, so dass Ereignisse, Gesichter und Aktionen der Hauptakteure besonders berücksichtigt werden sollten. Ereignisse und Objekte werden durch spezielle Kameraoperationen verstärkt, wie beispielsweise Zoom- und Zeitlupeneffekte

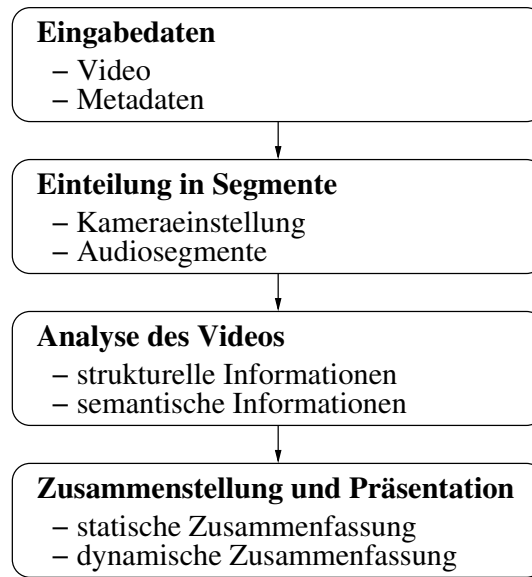


Abbildung 9.1: Erzeugung computergenerierter Zusammenfassungen von Videos

sowie vertikale Schwenks, die ein Objekt bzw. den Hintergrund hervorheben und als Merkmale zur Identifikation relevanter Kameraeinstellungen eingesetzt werden können [339, 392, 393]. Weitere wichtige semantische Informationen, aus denen besonders relevante Kameraeinstellungen für die Zusammenfassungen abgeleitet werden können, liefern Objekte im Bildvordergrund [7, 256]. Durch Analyse der Kamera- und Objektbewegungen wird zusätzlich ein Maß zur Beschreibung der *visuellen Komplexität* einer Kameraeinstellung abgeleitet. Anhand der Komplexität kann beispielsweise die minimale Zeit abgeschätzt werden, die notwendig ist, um den Inhalt einer Kameraeinstellung aufzunehmen [484, 485, 487].

Ein in mehreren Ansätzen berücksichtigtes Merkmal ist die *Bewegungsaktivität*, die indirekt das Tempo eines Videos beschreibt [77, 115, 482]. Es wird die Annahme getroffen, dass der visuelle Inhalt einer Kameraeinstellung mit geringer Bewegungsaktivität nur wenig variiert. Durch schnelles Abspielen der Kameraeinstellungen mit geringer Bewegungsaktivität wird die Dauer der Wiedergabe verkürzt. Der *Motion-Activity-Deskriptor* des MPEG-7-Standards kann aus der durchschnittlichen Länge der Bewegungsvektoren eines MPEG-Videos bzw. dessen Standardabweichung abgeleitet werden [235]. Ein weiterer Ansatz analysiert das Nutzerverhalten, um die Komplexität eines Videosegmentes zu bestimmen. Speziell für Lehrvideos und aufgezeichnete Präsentationen lässt sich anhand der Interaktion der Anwender auf besonders relevante oder schwer verständliche Bereiche des Videos schließen, die in einer Zusammenfassung kombiniert werden sollten [576].

Neben den automatisch berechneten Merkmalen zur Klassifikation relevanter Kameraeinstel-

lungen werden in mehreren Ansätzen manuell eingegebene *Metadaten* genutzt, um Zusammenfassungen zu erzeugen [405, 510, 511]. Spezielle Anwendungen unterstützen die Eingabe der manuellen Beschreibungen [352, 368, 374]. Zusätzlich kann ein Anwender die Auswahl der Kameraeinstellungen einer Zusammenfassung beeinflussen, indem spezielle Präferenzen wie beispielsweise die Länge der Zusammenfassung, der Anteil der Sprache oder die Stärke der Bewegungen berücksichtigt werden [86, 412, 459].

9.1.2 Genrespezifische Merkmale zur Auswahl von Kameraeinstellungen

Abhängig vom *Genre eines Videos* sind unterschiedliche Algorithmen zur Bewertung der Relevanz und Auswahl von Kameraeinstellungen geeignet. Die Verfahren unterscheiden sich durch die Art und Gewichtung der Merkmale, die in einem Video identifiziert werden. Eine automatische Erkennung des Genres eines Videos ist anhand der durchschnittlichen Länge der Kameraeinstellungen, der Farbgebung, der Bewegung und der Helligkeit möglich [142, 426, 427]. Im Folgenden werden für unterschiedliche Arten von Videos wesentliche Merkmale vorgestellt, die zur Auswahl der Kameraeinstellungen für eine Zusammenfassung geeignet sind.

In *Spielfilmen* sind Hauptdarsteller, schnelle Aktionen und besondere Ereignisse wie Explosionen oder plötzliche Lautstärkeänderungen besonders wichtig [323, 378]. Szenen liefern Informationen über zusammenhängende Kameraeinstellungen, Dialoge beschreiben Beziehungen zwischen den Personen im Video [5, 325, 417, 572]. Eines der ersten Systeme, das automatisch Merkmale eines Videos analysiert und eine computergenerierte Zusammenfassung erzeugt, ist das im Rahmen des Mannheimer Projektes *Movie-Content-Analysis* entwickelte *VAbstract* [323, 418, 419].

In *Sportveranstaltungen* sind besondere Ereignisse wie Tore, Strafstöße, Torschüsse oder Fouls für einen Zuschauer von besonderem Interesse [138, 139, 209]. Durch einfache Heuristiken können Aufnahmen in Zeitlupe, spezielle Frequenzen im Audiosignal zur Erkennung der Pfeife des Schiedsrichters, Lautstärkeänderungen durch den Jubel der Zuschauer oder Spielfeldmarkierungen erkannt werden. Diese Ereignisse liefern Hinweise auf interessante Segmente des Videos, die zu einer Zusammenfassung kombiniert werden [125, 425, 500, 535].

Nachrichtensendungen eignen sich durch die ausgeprägte Struktur besonders gut, um in kompakter Form eine Übersicht des Videos zu geben [83, 220, 463, 464, 465]. Algorithmen zur Erkennung von Texteinblendungen, zur Spracherkennung und zur Gesichtserkennung liefern wegen der qualitativ hochwertigen Studioaufnahmen häufig sehr zuverlässige Klassifikations-

ergebnisse. Zudem ist durch die große Überlappung der Themen eine Kombination mehrerer Nachrichtensendungen aus unterschiedlichen Sendern möglich [211]. Für Nachrichtensendungen bietet sich auch eine alternative Darstellung in Form einer Kollage an, in die geographische oder zeitliche Informationen eingeblendet werden können [84, 381, 533].

Neben Nachrichtensendungen zeichnen sich auch *Serien* durch einen sehr strukturierten Ablauf aus. Wegen der geringen Anzahl unterschiedlicher Orte und Personen ist die Komplexität beschränkt, so dass Verknüpfungen unter den einzelnen Kameraeinstellungen ermittelt und für eine Zusammenfassung berücksichtigt werden können [246]. Die Analyse mehrerer Folgen einer Serie ermöglicht es, eine Zusammenfassung aus mehreren Videos zu erstellen [557, 558, 559].

Bei Zusammenfassungen von *Amateurvideos* wie beispielsweise *Urlaubsvideos* soll ein Überblick über einen längeren Zeitraum gegeben werden. Die Analyse des Datums und der Uhrzeit der Aufnahme stellt sicher, dass in der Zusammenfassung Inhalte von unterschiedlichen Zeitpunkten enthalten sind [314, 316]. Anhand der Aufnahmezeit ist eine hierarchische Gruppierung der Kameraeinstellungen möglich. Die Aufbereitung der Urlaubsvideos kann unterstützt werden, indem beispielsweise automatisch eine passende Musik für das Video ausgewählt und unterlegt wird [215].

Bei der Erzeugung einer Zusammenfassung eines *Musikvideos* liegt der Schwerpunkt in der Analyse der Audiospur. Der Refrain eines Liedes ist besonders wichtig und sollte in der Zusammenfassung enthalten sein, wobei Gesichtsaufnahmen des Sängers häufig geeignete Bilder für die Zusammenfassung liefern [554, 555]. Weitere sehr spezialisierte Verfahren zur Erzeugung von Zusammenfassungen sind für *medizinische Videos* [337], sowie *Vorträge* und *Präsentationen* entwickelt worden [200].

9.1.3 Statische Zusammenfassungen von Videos

Die meisten Ansätze der in der Literatur vorgestellten Verfahren zur kompakten Darstellung eines Videos konzentrieren sich auf *statische Zusammenfassungen*, in denen einzelne aussagekräftige Bilder beispielsweise innerhalb einer Webseite angezeigt werden [184]. Die Bilder können durch zusätzliche Informationen in Form von textuellen Beschreibungen ergänzt werden. Die einfachste Form der Darstellung von Einzelbildern erfolgt als Liste oder Tabelle [67]. Hierbei wird für jede Kameraeinstellung ein repräsentatives Bild ausgewählt, für Kameraeinstellungen mit deutlichen Änderungen gegebenenfalls auch mehrere [74]. Bei längeren Videos ergibt sich eine umfangreiche Liste mit zum Teil mehreren tausend Einzelbildern,

die wegen der beschränkten Bildschirmgröße nicht mehr sinnvoll dargestellt werden können. Durch Gruppierung ähnlicher Bilder wird die Anzahl auf ein übersichtliches Maß reduziert [112, 121, 188, 331, 525].

Die in Kapitel 2.2 vorgestellten Distanzmaße zur Erkennung harter Schnitte sind geeignet, um Ähnlichkeiten zwischen Bildern zu erkennen [15]. Verschiedene Verfahren zur Gruppierung der Bilder, wie beispielsweise der *K-Means*-Algorithmus [135], die Analyse der *Korrelationsmatrix* [89, 90] oder die *Singulärwertzerlegung* (engl. *singular value decomposition*) [169], können zur Erzeugung statischer Zusammenfassungen eingesetzt werden [166, 168, 590, 591]. Um einen guten Überblick über das gesamte Video zu geben, werden möglichst unterschiedliche Bilder ausgewählt [479]. Zudem ist es möglich, anhand der gruppierten Bilder Beziehungen zwischen den einzelnen Kameraeinstellungen abzuleiten, um strukturelle Informationen des Videos zu erkennen [135, 309]. Als besonders relevant klassifizierte Kameraeinstellungen können mit Hilfe größerer Bilder hervorgehoben werden [48, 517, 518]. Durch die Gruppierung gehen jedoch Informationen über die zeitliche Struktur des Videos verloren.

Kamerabewegungen bleiben bei allen bisher vorgestellten Verfahren unberücksichtigt. Anstatt Einzelbilder des Videos in der statischen Zusammenfassung zu verwenden, werden in mehreren Ansätzen Hintergrund- bzw. Panoramabilder aus einer Kameraeinstellung erzeugt, so dass auch bei einem Kameraschwenk der komplette Bildhintergrund sichtbar ist [10, 11, 12, 153, 351, 494]. Eine weitere Möglichkeit zur Darstellung der Bildänderungen bieten dreidimensionale Volumenbilder [107]. Die Bewegung der Kamera wird vergleichbar mit Abbildung 2.3 in diesen Volumenbildern wiedergegeben, der Inhalt der Kameraeinstellung ist jedoch nur schwer zu erkennen.

Neben den Bildern können weitere Informationen für jede Kameraeinstellung hervorgehoben werden [322]. Wesentliche semantische Informationen wie Gesichter, Kameraeinstellungen mit vielen Veränderungen oder besondere Ereignisse beispielsweise in Sportsendungen lassen sich in den Bildern einer statischen Zusammenfassung durch Symbole oder Markierungen im Bild verdeutlichen.

9.1.4 Dynamische Zusammenfassungen von Videos

Bei einer *dynamischen Zusammenfassung* erfolgt die Darstellung in Form eines Videos, das die wesentlichen Inhalte in verkürzter Zeit wiedergibt. Dynamische Zusammenfassungen bieten den Vorteil, dass ein Wechsel des Mediums nicht erforderlich ist und sowohl Audio als auch bewegte Bilder verfügbar sind. Ein sehr einfaches Verfahren erzeugt dynamische Zusam-

menfassungen durch eine Erhöhung der Bildwiederholrate [250, 545].

Falls eine Zusammenfassung auch Audio enthalten soll, ist eine sinnvolle Erhöhung der Bildwiederholrate um bis zu 60 Prozent möglich [377]. Dabei muss verhindert werden, dass sich die Tonhöhe durch das schnellere Abspielen der Audiospur verändert. Im digitalen Audiostrom ist für jedes Zeitintervall definiert, wann und wie lange eine bestimmte Frequenz wiedergegeben wird. Durch eine Verkürzung der Länge dieses Zeitfensters wird die Abspieldauer entsprechend gekürzt [8, 198, 199]. Bei starken Änderungen gleicht ein Glättungsfilter ein mögliches Klicken und Verzerrungen im Bereich der Übergänge aus [399].

Dynamische Zusammenfassungen können durch spezielle Interaktionsmöglichkeiten erweitert werden, so dass der Inhalt des Videos noch schneller erfasst werden kann [490]. Der Anwender hat die Möglichkeit, die Abspielgeschwindigkeit durch Schieberegler zu verändern [221], semantisch zusammenhängende Videosegmente zu überspringen [573] und individuelle Abspielpräferenzen festzulegen [306].

Zu den Systemen und Projekten, die computergenerierte Zusammenfassungen oder eine effiziente Navigation innerhalb eines Videos ermöglichen, zählen *CueVideo* [421], *Informedia* [532, 534], *MoCA* [419] und das *Hitchcock*-System zum semiautomatischen Editieren von Videos [161, 162]. Das *Informedia*-Projekt hat zwei Anwendungen entwickelt, um die Inhalte von Nachrichtensendungen leichter zugänglich zu machen [533]. Die erste Anwendung stellt eine Oberfläche zur Navigation und Präsentation von Zusammenfassungen zur Verfügung [85]. Relevante Wörter werden durch Spracherkennungsalgorithmen, Texte und Gesichter durch Bildanalyseverfahren identifiziert. Bei der zweiten Anwendung erfolgt die Darstellung des Videos in Form einer Kollage, in der die Inhalte mehrerer Nachrichtenvideos gleichzeitig dargestellt werden [84, 381].

Das *CueVideo*-System ermöglicht es, ein Video durch Erhöhung der Bildwiederholrate schneller abzuspielen, und passt das Audiosignal unter Beibehaltung der Tonhöhe an [8, 399]. Im Rahmen des *MoCA*-Projektes (*Movie Content Analysis*) ist eine der ersten Anwendungen zur automatischen Erzeugung von dynamischen Zusammenfassungen entstanden [323]. Das System analysiert Spielfilme und identifiziert spezielle Ereignisse wie Explosionen, Pistolenschüsse oder Dialoge, die in die Zusammenfassung übernommen werden.

9.2 Systemüberblick

Im Rahmen des in Kapitel 2.3.6 vorgestellten Projektes *European Chronicles Online* wurde ein Videoarchiv für historische Dokumentationen entwickelt. Teil des Systems ist eine von

uns entwickelte Komponente, die automatisch Zusammenfassungen von Videos erzeugt und diese den Anwendern als zusätzliche Darstellungsmöglichkeit des Videos zur Verfügung stellt. Im Folgenden werden unsere neuen Algorithmen zur automatischen Erzeugung von Zusammenfassungen vorgestellt. Trotz der Vielzahl an bestehenden Verfahren hat jeder Ansatz spezifische Schwächen und ist zur Erzeugung von Zusammenfassungen für historische Video-Dokumentationen nur bedingt geeignet.

Der erste Schritt umfasst die Analyse des Videos, um relevante Merkmale in einzelnen Kameraeinstellungen zu ermitteln. Neue Algorithmen zur Beschreibung der strukturellen und semantischen Informationen eines Videos werden in diesem Zusammenhang vorgestellt. Anschließend erfolgt die Bewertung, Auswahl und Kombination der einzelnen Kameraeinstellungen. Hierzu wird ein neues heuristisches, iteratives Verfahren vorgeschlagen, das allgemein einsetzbar und nicht nur auf historische Dokumentationen beschränkt ist. Zusätzlich werden noch Verbesserungen speziell für historische Dokumentationen vorgeschlagen, um beispielsweise zu verhindern, dass Kameraeinstellungen von sehr schlechter Bildqualität in die Zusammenfassung aufgenommen werden. Im Gegensatz zu statischen Zusammenfassungen, bei denen die Darstellung durch einzelne repräsentative Bilder erfolgt, werden im Falle von dynamischen Zusammenfassungen Kameraeinstellungen miteinander kombiniert und nach Anpassung der Audiospur als Video gespeichert.

Abbildung 9.2 verdeutlicht die wesentlichen Schritte bei der Erzeugung einer Zusammenfassung. Bei der Analyse der Struktur des Videos werden neben Kameraeinstellungen, Szenen und Dialogen auch Gruppen von Kameraeinstellungen mit ähnlichen Bildinhalten identifiziert. Im Unterschied zu einer Szene enthalten die Kameraeinstellungen innerhalb einer Gruppe keinen zeitlichen Bezug und können über das ganze Video verteilt sein.

Nach dem Analyseschritt werden die Merkmale zur Beschreibung der Kameraeinstellungen gewichtet und einzelne Kameraeinstellungen für die Zusammenfassung ausgewählt und angeordnet. Bei einer statischen Zusammenfassung werden repräsentative Bilder für die ausgewählten Kameraeinstellungen gespeichert. Für eine dynamische Zusammenfassung werden die relevanten Kameraeinstellungen mit der Audiospur neu zu einem Video kombiniert.

9.3 Strukturelle und semantische Analyse des Videos

Bei der automatischen Analyse des Videos werden Informationen über Schnitte, Kamerabewegungen, Gesichter, Objekte und Textregionen ermittelt. Um Kameraeinstellungen zu bewerten, werden Informationen auf der Ebene der Kameraeinstellungen zusammengefasst und

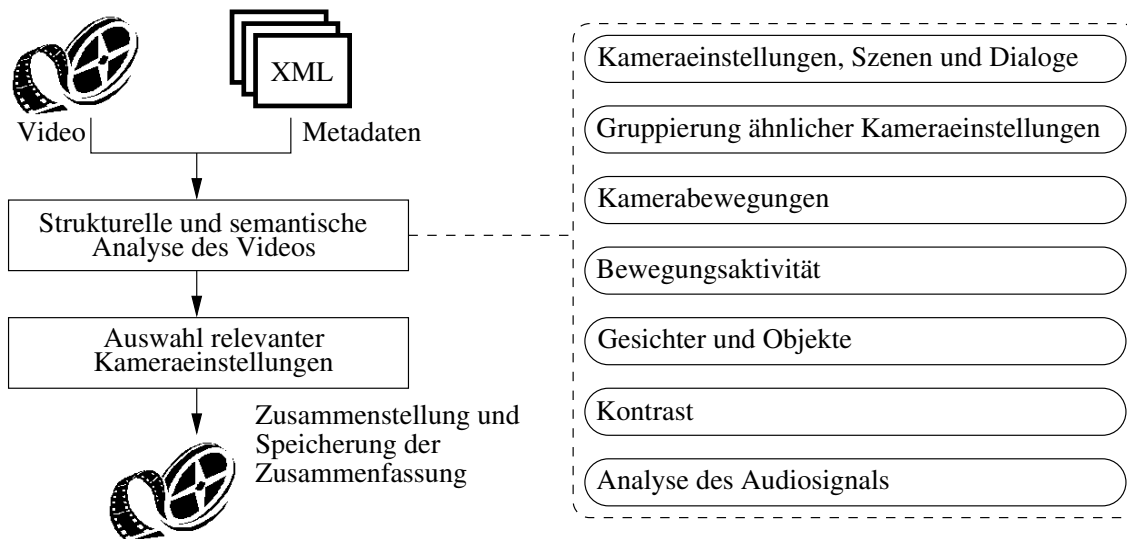


Abbildung 9.2: Überblick zur Erzeugung von Zusammenfassungen

durch einen aggregierten Merkmalswert beschrieben.

Algorithmen zur Berechnung und Aggregation der strukturellen und semantischen Informationen werden im Folgenden vorgestellt. Das Auffinden repräsentativer Einzelbilder erfolgt mit Hilfe eines neuen Algorithmus, bei dem fehlerhafte Bilder in historischen Videos identifiziert und für die Zusammenfassung ausgeschlossen werden. Anschließend schlagen wir einen neuen Algorithmus zur effizienten Gruppierung ähnlicher Kameraeinstellungen vor, der auch die Erkennung von Kameraeinstellungen in schlechter Bildqualität ermöglicht. Als weiteres semantisches Merkmal wird ein Maß für die Bewegungsaktivität vorgestellt, das sowohl plötzlich auftretende Pixeländerungen als auch starke Bewegungen berücksichtigt.

9.3.1 Schnitterkennung und Auswahl repräsentativer Einzelbilder

Die Erkennung der Schnitte in den historischen Videos erfolgt mit dem in Kapitel 2.3.6 vorgestellten Verfahren. Harte Schnitte sowie Ein-, Aus- und Überblendungen werden erkannt, wobei der Anteil der weichen Schnitte in den historischen Videos sehr gering ist, da die manuelle Erzeugung mit einem hohen Aufwand verbunden war.

Für eine statische Zusammenfassung und zur Erkennung von ähnlichen Kameraeinstellungen werden *repräsentative Bilder* von allen Kameraeinstellungen benötigt. Zur Ermittlung der Bilder schlagen wir das folgende neue Verfahren vor: Zunächst wird das mittlere Bild einer Kameraeinstellung als repräsentatives Bild ausgewählt. In den historischen Videos treten häufig fehlerhafte Bildbereiche und zum Teil vollständig defekte Bilder auf. Durch einen Vergleich

des Histogramms des festgelegten Bildes mit dem durchschnittlichen Histogramm aller Bilder der Kameraeinstellung kann verhindert werden, dass einzelne fehlerhafte Bilder verwendet werden. Bei einer großen Differenz beider Histogramme wird das repräsentative Bild durch das Bild der Kameraeinstellung ersetzt, dessen Histogramm möglichst ähnlich dem durchschnittlichen Histogramm ist. Die Qualität der ausgewählten Bilder steigt deutlich, da in den analysierten Videos nur sehr selten fehlerhafte Bildbereiche während der gesamten Kameraeinstellung auftreten.

9.3.2 Gruppierung ähnlicher Kameraeinstellungen

Die ausgewählten repräsentativen Bilder werden verwendet, um ähnliche Kameraeinstellungen zu identifizieren und zu Gruppen zu aggregieren. Die *Größe einer Gruppe* wird als Summe der Länge der Kameraeinstellungen dieser Gruppe definiert und gibt einen Hinweis auf die Bedeutung der Gruppe für das Video. Bei der Auswahl der einzelnen Kameraeinstellungen erhalten besonders große Gruppen eine hohe Priorität, so dass diese Gruppen durch mindestens eine Kameraeinstellung in der Zusammenfassung repräsentiert werden. Im Unterschied zu Szenen, bei denen es sich um eine semantische Gruppierung von benachbarten Kameraeinstellungen handelt, ist ein zeitlicher Bezug innerhalb einer Gruppe mit ähnlichen Kameraeinstellungen nicht erforderlich.

Die Zuordnung zu Gruppen erfolgt durch einen Vergleich der repräsentativen Bilder. Graustufenhistogramme von neun gleich großen Bildregionen werden als Merkmalsvektor eingesetzt, um die Ähnlichkeit zwischen Bildern zu bestimmen. Die Summe der absoluten Differenzen wird als Differenzmaß für die Histogramme verwendet.

Der neue Algorithmus zur *Bildung der Gruppen* wird im Folgenden erläutert: Zunächst werden spezielle Zentren für jede Gruppe identifiziert. Sowohl die repräsentativen Bilder als auch die Zentren werden durch Graustufenhistogramme abgebildet und beschreiben jeweils einen Punkt in einem mehrdimensionalen Raum. Während der Gruppierung werden neue Zentren festgelegt, bis der Abstand aller Bilder zum jeweils nächstgelegenen Zentrum einen Schwellwert unterschreitet. Falls der Abstand mindestens eines repräsentativen Bildes über dem Schwellwert liegt, wird ein zusätzliches Zentrum benötigt und hinzugefügt. Folgender von uns neu entwickelter Algorithmus wird zur *Bildung von Gruppen mit ähnlichen Kameraeinstellungen* eingesetzt:

1. Das erste Zentrum wird als durchschnittliches Histogramm aller repräsentativen Bilder initialisiert. Die Summe der Abstände zwischen dem Zentrum und allen Bildern ist für

diesen Punkt minimal.

2. Für jedes repräsentative Bild wird das nächstgelegene Zentrum identifiziert, wobei direkt nach der Initialisierung nur ein Zentrum existiert. Jedes Bild wird dem nächstgelegenen Zentrum zugeordnet, und der Abstand zwischen Zentrum und Bild wird berechnet.
3. Die Positionen aller Zentren werden aktualisiert. Die neue Position eines Zentrums ist definiert als durchschnittlicher Histogrammwert aller Bilder, die diesem Zentrum zugeordnet sind.
4. Das Bild mit dem größten Abstand zu seinem Zentrum wird ausgewählt. Falls der Abstand über einem Schwellwert liegt, sind die Unterschiede innerhalb der Gruppe sehr hoch, und ein neues Zentrum wird an der Position dieses Bildes eingefügt. Der Algorithmus wird mit Schritt 2 fortgesetzt, bis alle repräsentativen Bilder innerhalb einer Gruppe eine starke Ähnlichkeit besitzen.

In sehr kurzen Videos mit wenigen Kameraeinstellungen ist es möglich, dass die Anzahl der Gruppen und Kameraeinstellungen einander entsprechen. In Serien, Nachrichtensendungen und Sportveranstaltungen gibt es im Allgemeinen sehr große Gruppen mit vielen Kameraeinstellungen.

Der Algorithmus zur Gruppierung von Kameraeinstellungen kann erweitert werden, um fehlerhafte Kameraeinstellungen zu identifizieren. In vielen historischen Videos des *European-Chronicles-Online*-Systems sind sowohl einzelne Bilder als auch längere Segmente des Videos beschädigt; diese sollten in einer Zusammenfassung nicht enthalten sein. Zur Identifikation fehlerhafter Kameraeinstellungen werden einzelne Zentren festgelegt, die *auf keinen Fall* in der Zusammenfassung enthalten sein sollen. Die Gruppen mit den vordefinierten Zentren enthalten fehlerhafte oder qualitativ geringwertige Kameraeinstellungen und werden als *defekte Gruppen* bezeichnet. Wird ein Bild einer defekten Gruppe zugeordnet, so bleibt die entsprechende Kameraeinstellung für die Zusammenfassung unberücksichtigt. Abbildung 9.3 zeigt beispielhaft drei Gruppen, in denen jeweils die repräsentativen Bilder zweier Kameraeinstellungen enthalten sind. Rechts ist eine defekte Gruppe abgebildet, die durch ein sehr dunkles Bild initialisiert wurde.

9.3.3 Erkennung von Szenen

Eine *Szene* besteht aus mehreren benachbarten Kameraeinstellungen, die eine zusammenhängende Handlung beschreiben. Im Gegensatz zu eine Gruppe mit ähnlichen Kameraeinstellun-

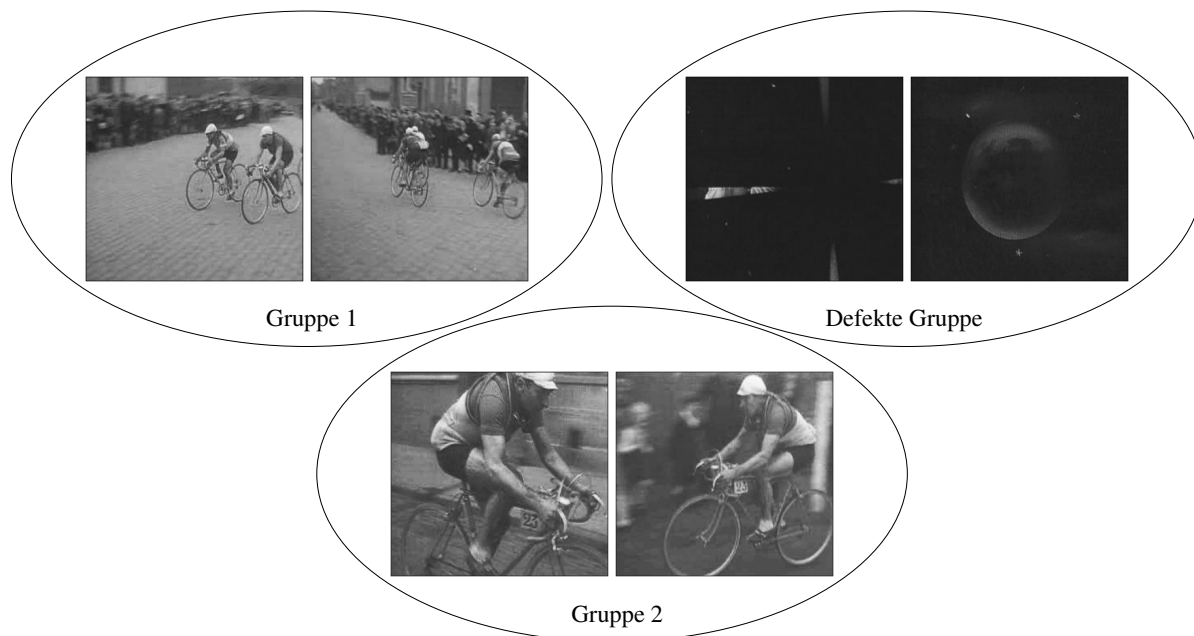


Abbildung 9.3: Ergebnisse des Algorithmus zur Gruppierung ähnlicher Kameraeinstellungen: Die rechte Gruppe enthält ein vordefiniertes Zentrum, das durch ein nahezu schwarzes Bild definiert ist. Kameraeinstellungen dieser Gruppe werden für die Zusammenfassung nicht berücksichtigt.

en handelt es sich um eine semantische Einheit des Videos. Üblicherweise spielt eine Szene an einem Ort, so dass der Bildhintergrund in allen Kameraeinstellungen eine hohe Übereinstimmung aufweist. Auch ein Schwenk der Kamera oder die Aufnahme aus einer anderen Blickrichtung verursachen im Allgemeinen nur geringe Veränderungen des Bildhintergrundes. Zur Erkennung der einzelnen Szenen werden die Gruppen mit ähnlichen Kameraeinstellungen analysiert. Eine Szene besteht aus benachbarten Kameraeinstellungen und soll maximal zwei Gruppen zugeordnet sein. Mit Hilfe der in Kapitel 7 vorgestellten Algorithmen zur Gesichtserkennung werden *Dialoge* als spezielle Ausprägung einer Szene identifiziert, bei der die Kamera zwischen zwei oder mehreren Personen wechselt.

9.3.4 Kamerabewegung

Bewegungen zählen zu den wichtigsten Merkmalen eines Videos. Dabei ist insbesondere die semantische Beschreibung der Kamerabewegung innerhalb einer Kameraeinstellung und nicht die exakte Beschreibung des Kameramodells (vgl. Gleichung 3.1) zwischen zwei benachbarten Bildern wichtig. Durch eine Aggregation der Kamerabewegung über mehrere Bilder können Schwenks, Zoomeffekte und Rotationen identifiziert werden. Bei der Bewertung werden nur

deutliche Kamerabewegungen berücksichtigt, und verwackelte Aufnahmen bleiben unberücksichtigt.

Kamerabewegungen und Kameraoperationen geben Hinweise auf besonders wichtige Segmente des Videos. Bei einem eingehenden Zoomeffekt ist häufig das Objekt im Bildzentrum von zentraler Bedeutung. Wie auch die Analyse der Kameraoperationen in Kapitel 3.5 gezeigt hat, werden vertikale Schwenks sehr selten eingesetzt und lenken die Aufmerksamkeit auf die Umgebung bzw. den Bildhintergrund. In Ausnahmefällen ist es möglich, dass eine Kameraeinstellung mehr als eine deutlich ausgeprägte Kamerabewegung enthält. Die Kameraeinstellung wird dann in mehrere Segmente unterteilt, die unabhängig voneinander analysiert werden.

9.3.5 Bewegungsaktivität

Ein weiteres wichtiges Merkmal zur Beschreibung von Kameraeinstellungen ist die *Bewegungsaktivität*. Im Rahmen des Auswahlprozesses wird angenommen, dass Kameraeinstellungen mit starken Bewegungen besonders wichtig sind, da mehrere unterschiedliche Bildinhalte pro Zeitintervall gezeigt werden. Eine deutliche Änderung zwischen zwei benachbarten Bildern innerhalb einer Kameraeinstellung kann auf eine schnelle Kamerabewegung, eine Objektbewegung eines großen Objektes oder auf besondere Ereignisse wie beispielsweise Lichtänderungen, Feuer oder Explosionen zurückgeführt werden.

Ein aggregierter Wert zur Beschreibung der Bewegungsaktivität wird für jede Kameraeinstellung berechnet. Hierzu werden zwei Maße zur Beurteilung der Bewegungsaktivität kombiniert. Das erste Maß leitet sich aus der Summe der absoluten Pixeldifferenzen zweier benachbarter Bilder ab. Das zweite Maß analysiert die durchschnittliche Länge der Bewegungsvektoren und ist vergleichbar mit dem *Motion-Activity-Deskriptor* des MPEG-7 Standards. Beide Werte werden gleich gewichtet und zu einem aggregierten Wert zusammengefasst, der die Bewegungsaktivität der Kameraeinstellung beschreibt. Die Korrelation zwischen beiden Maßen ist sehr hoch, wobei Helligkeitsänderungen durch Feuer oder Explosionen besonders im ersten Maß und schnelle Kamerabewegungen stärker im zweiten Maß berücksichtigt werden.

9.3.6 Gesichter und Objekte

Große *Gesichter* oder *Objekte* im Bildzentrum einer Kameraeinstellung haben in Dokumentationen häufig eine besondere Bedeutung. Im Gegensatz zu Spielfilmen, in denen Hauptdarsteller in Nahaufnahme gezeigt werden, sind in den historischen Dokumentationen häufig bekannte Persönlichkeiten wie beispielsweise herausragende Sportler, Wissenschaftler oder Politiker

zu sehen. Kameraeinstellungen mit großen Gesichtern gelten als besonders wichtig und sollten in der Zusammenfassung enthalten sein.

Objekte liefern weitere wichtige semantische Informationen über ein Video. Wird ein Objekt besonders häufig im Video erkannt, so sollte es auch in der Zusammenfassung erscheinen. Insbesondere für Sportereignisse, in denen einzelne Personen oder Fahrzeuge wiederholt im Bild sichtbar sind, liefert diese Heuristik eine gute Auswahl an Kameraeinstellungen.

9.3.7 Analyse des Audiosignals

Ein Betrachter empfindet es als sehr unangenehm, wenn der Ton mitten in einem Satz oder in besonders lauten Abschnitten unterbrochen wird. Um geeignete Bereiche für einen Schnitt der Audiospur zu finden, werden ruhige Segmente identifiziert. Ein Bereich gilt als ruhig, falls die Energie des Audiosignals für die Dauer von mindestens einer Sekunde unter einem Schwellwert liegt.

Die Qualität des Audiosignals variiert in den analysierten historischen Videos sehr stark. Mit Ausnahme der Stummfilme sind viele Videos mit Musik unterlegt und enthalten Rauschen und deutliche Hintergrundgeräusche, die beispielsweise durch den zum Teil fast einhundert Jahre alten Filmprojektor entstanden sind. Vor der Analyse der Audiospur erfolgt eine Normierung anhand der maximalen Lautstärke des Videos.

9.4 Auswahl relevanter Kameraeinstellungen

In diesem Abschnitt wird eine neue Heuristik zur Auswahl geeigneter Kameraeinstellungen vorgestellt. Abbildung 9.4 verdeutlicht die wesentlichen Schritte bei deren Auswahl. Durch den hohen Anteil fehlerhafter Kameraeinstellungen in historischen Videos werden zunächst Kameraeinstellungen ausgeschlossen, die auf keinen Fall in der Zusammenfassung erscheinen sollen. Auswahlkriterien sind ein sehr geringer Kontrast oder die Zuordnung zu einer defekten Gruppe. Auch sehr kurze Kameraeinstellungen mit einer Länge von weniger als drei Sekunden werden nicht für eine Zusammenfassung ausgewählt.

Um die berechneten strukturellen und semantischen Informationen miteinander vergleichen zu können, werden *aggregierte Merkmalswerte* berechnet, welche diese Informationen auf einen Wert im Intervall $[0, 1]$ abbilden und eine Bewertung von Kameraeinstellungen ermöglichen. Tabelle 9.1 beschreibt die Merkmale, die zur Auswahl der Kameraeinstellungen berücksichtigt werden. Der größte Teil der aggregierten Merkmalswerte wird nur einmal berechnet und ändert

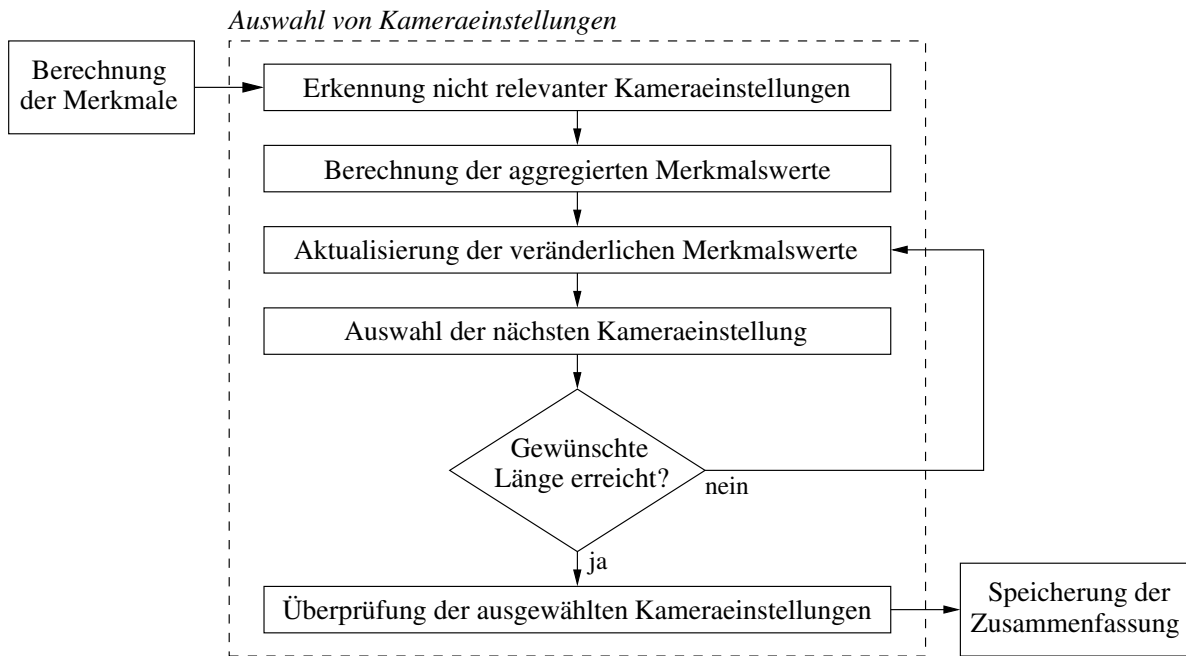


Abbildung 9.4: Auswahl von Kameraeinstellungen

sich während der Auswahl der Kameraeinstellungen nicht. Drei Merkmale, die als *veränderliche Merkmale* bezeichnet werden, müssen nach jeder neu ausgewählten Kameraeinstellung aktualisiert werden.

Im Rahmen der Bewertung der semantischen Informationen werden neue Heuristiken vorgeschlagen, um aus der Beschreibung eines Merkmals einen Wert zur Beurteilung der Relevanz einer Kameraeinstellung für eine Zusammenfassung abzuleiten. In die Heuristik zur Bewertung der Kamerabewegung gehen beispielsweise die Art, Stärke und Dauer einer Kameraoperation ein. Die Auswahl der Kameraeinstellungen erfolgt in einem iterativen Prozess, wobei bereits ausgewählte Kameraeinstellungen den weiteren Auswahlprozess durch die veränderlichen Merkmalswerte beeinflussen. Neue Heuristiken zur Bewertung der Ähnlichkeit und zur Verteilung der Kameraeinstellungen über die gesamte Länge des Videos werden in diesem Zusammenhang vorgeschlagen. Die Algorithmen zur Berechnung der einzelnen aggregierten Merkmalswerte der Tabelle 9.1 werden im Folgenden näher erläutert.

9.4.1 Bewertung der Kamerabewegung

Kamerabewegungen können dazu eingesetzt werden, um besondere Inhalte des Videos hervorzuheben. Ausgehende Zoomeffekte und Schwenks liefern Informationen über den Bildhintergrund bzw. den Ort der Handlung. Bei einem eingehenden Zoomeffekt wird das Zentrum des

Merkmale	Verfügbare Informationen	Zeitintervall	veränderliches Merkmal
Kamera-bewegung	Art der Kamerabewegung (Zoom, Schwenk), Stärke der Bewegung	Teil einer Kameraeinstellung	nein
Bewegungs-aktivität	Umfang der Bewegungsaktivität	Bild	nein
Gesicht	Größe, Position, Rotationswinkel	Bild	nein
Objekt	Größe, Objektname, Name der Objektklasse, Zuverlässigkeit	Bild	nein
Kontrast	Kontrast eines Bildes	Bild	nein
Gruppen ähnlicher Kameraeinstellungen	Liste mit Kameraeinstellungen	Kameraeinstellung	ja
Szene	Liste mit Kameraeinstellungen	Kameraeinstellung	ja
Zeitliche Verteilung	Entfernung zur nächsten ausgewählten Kameraeinstellung	Kameraeinstellung	ja
Audio	Zeitintervalle der ruhigen Bereiche	Teil des Videos	nein

Tabelle 9.1: Aggregierte Merkmale zur Beschreibung der Kameraeinstellungen

Bildes hervorgehoben, in dem beispielsweise ein besonderes Objekt oder eine für das Video relevante Person abgebildet ist. Der aggregierte Wert zur Beschreibung der Kamerabewegung C_A wird durch die *Art* der Bewegung C_T , die *Stärke* der Kamerabewegung C_S und deren *Dauer* C_L beeinflusst:

$$C_A = \min (C_T + C_S + C_L, 1) \quad \text{mit} \quad (9.1)$$

$$C_S = \min (T_S \cdot V_{MV}, 0,5) \quad (9.2)$$

$$C_L = \min (T_L \cdot V_L, 0,5) \quad (9.3)$$

Abhängig von der Art der Kamerabewegung sind unterschiedliche Werte für C_T definiert. Die geringste Bedeutung haben horizontale Schwenks und ausgehende Zoomoperationen ($C_T = 0,2$). Selten treten vertikale Schwenks auf, die eine stärkere Gewichtung erhalten ($C_T = 0,3$). Die größte Bedeutung haben eingehende Zoomoperationen ($C_T = 0,4$), da sie häufig wichtige Objekte im Bildzentrum zeigen. Falls nach einer deutlichen Kamerabewegung die Kamera für mindestens zehn Sekunden statisch auf einem Bildausschnitt fokussiert bleibt, wird wegen der zu erwartenden besonderen Bedeutung dieser Kameraeinstellung der Wert von C_T zusätzlich

um 0,1 erhöht.

Die Stärke der Kamerabewegung C_S wird aus der durchschnittlichen Länge V_{MV} der Bewegungsvektoren des Kameramodells abgeleitet und mit dem Faktor T_S in Abhängigkeit der Bildbreite gewichtet. Der Skalierungsfaktor T_L gewichtet die Dauer der erkannten Kamerabewegung V_L , so dass der maximale Wert von 0,5 bei starken Kamerabewegungen ab einer Länge von zehn Sekunden erreicht wird.

9.4.2 Bewertung der Bewegungsaktivität

Die *Bewegungsaktivität* ist definiert als normierte Summe der beiden Aktivitätswerte, die aus der Bilddifferenz und der Länge der Bewegungsvektoren ermittelt werden. Der Durchschnitt aller Bilder einer Kameraeinstellung definiert den aggregierten Merkmalswert.

9.4.3 Bewertung der Gesichter und Objekte

Der *aggregierte Gesichtswert* wird aus dem Anteil der Gesichtspixel eines Bildes abgeleitet. Zwei Gesichter mittlerer Größe erhalten somit eine ähnliche Bedeutung wie ein großes Gesicht. Der durchschnittliche Wert aller Bilder einer Kameraeinstellung definiert den aggregierten Gesichtswert.

Mit Hilfe der in Kapitel 5 vorgestellten Algorithmen ist es grundsätzlich möglich, *Objekte* der Objektklassen *Flugzeug*, *Schiff*, *PKW* und *Person* automatisch in den Videos zu identifizieren. Im Vergleich zu aktuellen Videos haben Schiffe, Flugzeuge und PKWs in den analysierten historischen Dokumentationen, die Anfang bis Mitte des letzten Jahrhunderts aufgenommen wurden, eine wesentlich größere Bedeutung. Die Heuristik zur Berechnung des aggregierten Wertes berücksichtigt die Anzahl, Größe und Zuverlässigkeit der erkannten Objekte innerhalb einer Kameraeinstellung. Wird dieselbe Objektklasse mehrfach im Video identifiziert, so erhöhen sich die aggregierten Werte dieser Kameraeinstellungen zusätzlich.

Falls ein Objekt erkannt wird, kann zusätzlich eine Aussage über die Qualität einer Kameraeinstellung abgeleitet werden. Die Qualität der Kameraeinstellung, in denen Objekte erkannt werden, muss sehr hoch sein, da bei geringer Bildschärfe Fehler im Hintergrundbild entstehen und durch starkes Rauschen oder Bildfehler eine zuverlässige Segmentierung nicht möglich ist. Kameraeinstellungen mit erkannten Objekten werden besonders berücksichtigt, indem in der Heuristik zur Bewertung der Objekte die Untergrenze des aggregierten Wertes mit 0,5 festgelegt ist.

9.4.4 Bewertung des Kontrastes

In historischen Videos ist die Bildqualität zum Teil so schlecht, dass der Inhalt nur schwer oder gar nicht erkannt werden kann. Daher liefert der Kontrast eines Bildes einen guten Hinweis über die Bildqualität einer Kameraeinstellung, die im Falle eines besonders niedrigen Kontrastes nicht ausgewählt werden sollte. Der *aggregierte Kontrast* ist definiert als der durchschnittliche auf das Intervall $[0, 1]$ normierte Kontrast aller Bilder der Kameraeinstellung.

9.4.5 Bewertung der Ähnlichkeit von Kameraeinstellungen

Alle bisher beschriebenen aggregierten Merkmalswerte werden einmalig initialisiert und bleiben während der Auswahl der Kameraeinstellungen unverändert. Die aggregierten Werte zur Beschreibung von ähnlichen Szenen, Kameraeinstellungen und deren zeitlicher Verteilung werden durch jede neu ausgewählte Kameraeinstellung beeinflusst und müssen regelmäßig aktualisiert werden.

Kameraeinstellungen mit visueller Ähnlichkeit werden gemeinsamen Gruppen zugeordnet. Um einen möglichst umfangreichen Überblick in der Zusammenfassung zu geben, sollten Kameraeinstellungen aus unterschiedlichen Gruppen ausgewählt werden. Die Bewertung C_i einer Gruppe i hängt von dessen Länge ab, d. h. von der Summe der Längen aller Kameraeinstellungen, die dieser Gruppe zugeordnet sind:

$$C_i = \frac{1}{\max_j \{D_j\}} \cdot \frac{D_i}{1 + S_i^2}, \quad j = 1 \dots N. \quad (9.4)$$

D_i definiert die Länge der Gruppe i , S_i gibt die Anzahl der bereits ausgewählten Kameraeinstellungen dieser Gruppe an. Die größte Gruppe innerhalb des Videos definiert den Gewichtungsfaktor zur Normierung von C_i auf das Intervall $[0, 1]$. Mit der Auswahl einer Kameraeinstellung aus der Gruppe i erhöht sich S_i um eins, so dass für den weiteren Auswahlprozess der aggregierte Wert dieser Gruppe sinkt und bevorzugt Kameraeinstellungen aus anderen großen Gruppen ausgewählt werden. Alle Kameraeinstellungen der Gruppe i erhalten C_i als aggregierten Wert zur Bewertung der Ähnlichkeit zugewiesen.

9.4.6 Bewertung der Szenen

Damit der Inhalt einer Szene leichter verständlich ist und keine unpassenden Schnitte im Audiosignal entstehen, sollten benachbarte Kameraeinstellungen einer Szene in der Zusammenfassung enthalten sein. Dabei liefert eine einzelne Kameraeinstellung häufig nicht ausreichend

Informationen, um den Inhalt der Szene zu verstehen. Andererseits wiederholen sich bei mehr als zwei ausgewählten Kameraeinstellungen einer Szene die Inhalte, und der Zugewinn an Informationen nimmt deutlich ab.

Die Heuristik zur Bewertung der Szenen initialisiert den Wert für jede Kameraeinstellung zunächst mit 0,5. Falls zwei oder mehr Kameraeinstellungen einer Szene für die Zusammenfassung ausgewählt sind, wird mit jeder weiteren Kameraeinstellung der Wert um 20 Prozent reduziert. Damit möglichst zwei benachbarte Kameraeinstellungen für die Zusammenfassung ausgewählt werden, erhalten bei genau einer ausgewählten Kameraeinstellung die Werte der angrenzenden Kameraeinstellungen derselben Szene den Maximalwert von eins. Gleichzeitig werden die Werte der anderen Kameraeinstellungen dieser Szene auf null reduziert. Durch diese Heuristik wird die Auswahl von genau zwei benachbarten Kameraeinstellungen begünstigt. Mit den Bewertungen der Szenen und der ähnlichen Kameraeinstellungen werden unterschiedliche Ziele verfolgt. Um das Verständnis zu erleichtern, sollen aus einer Szene möglichst zwei benachbarte Kameraeinstellungen für die Zusammenfassung ausgewählt werden. Mit der Bildung der Gruppen wird das Ziel verfolgt, viele Kameraeinstellungen mit deutlichen visuellen Unterschieden in die Zusammenfassung aufzunehmen.

9.4.7 Bewertung der zeitlichen Verteilung

Innerhalb einer Zusammenfassung soll der gesamte Inhalt und nicht nur einzelne Teile des Videos gezeigt werden. Durch eine möglichst gute Verteilung der ausgewählten Kameraeinstellungen über die gesamte Länge des Videos kann dieses Ziel unterstützt werden. Eine gute zeitliche Verteilung ist besonders für Dokumentationen und Nachrichtensendungen wichtig, für die ein Überblick über das Video gegeben werden soll. Bei Spielfilmen muss diese Heuristik eingeschränkt werden, da in einer Vorschau beispielsweise das spannende Ende des Filmes nicht aufgedeckt werden soll. Ungeeignet ist die Heuristik zur Bewertung der zeitlichen Verteilung für Zusammenfassungen von Sportveranstaltungen, da besondere Aktionen und Ereignisse relevant sind, die nicht gleichmäßig über die gesamte Länge des Videos verteilt sind. Die Bewertung der *zeitlichen Verteilung* soll dazu führen, dass Kameraeinstellungen aus den unterschiedlichen Bereichen des Videos ausgewählt werden. Der aggregierte Wert wird aus dem Abstand der Kameraeinstellung zu der am nächsten gelegenen ausgewählten Kameraeinstellung abgeleitet und auf das Intervall $[0, 1]$ normiert. Abbildung 9.5 verdeutlicht beispielhaft die Berechnung des Wertes der zeitlichen Verteilung für eine Videosequenz mit bereits drei ausgewählten Kameraeinstellungen.

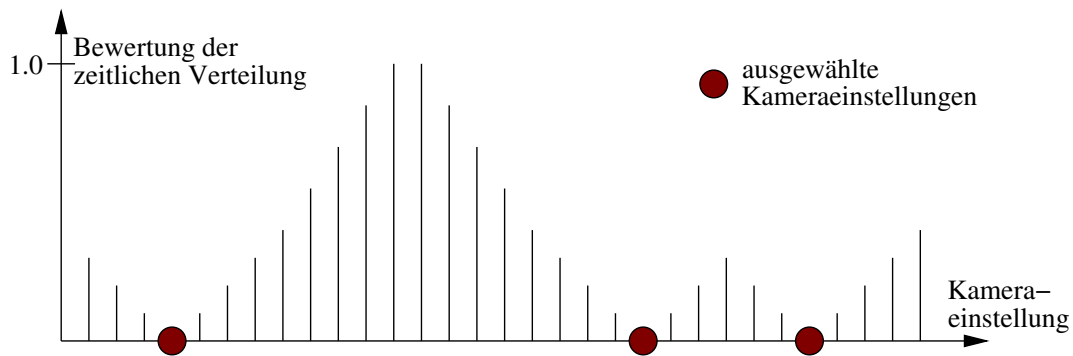


Abbildung 9.5: Schematische Darstellung der Berechnung des Wertes für die zeitliche Verteilung. Der Wert einer Kameraeinstellung steigt mit der Entfernung zur nächstgelegenen ausgewählten Kameraeinstellung.

9.5 Erzeugung einer Zusammenfassung

Nach der Berechnung der aggregierten Merkmalswerte erfolgt die Auswahl der Kameraeinstellungen für die Zusammenfassung. Bei den folgenden Überlegungen wird davon ausgegangen, dass eine dynamische Zusammenfassung erzeugt wird, in der die Audiospur und passende Übergänge zwischen den Kameraeinstellungen berücksichtigt werden. Vor der Speicherung der Zusammenfassung werden die ausgewählten Kameraeinstellungen anhand von Regeln überprüft, die die Qualität der Zusammenfassungen der historischen Dokumentationen signifikant verbessern. Um eine statische Zusammenfassung zu erhalten, kann für jede Kameraeinstellung der dynamischen Zusammenfassung ein repräsentatives Bild ausgewählt werden.

9.5.1 Auswahl von Kameraeinstellungen

Die Auswahl der Kameraeinstellungen erfolgt durch Analyse der aggregierten Merkmalswerte. Der *gewichtete Wert einer Kameraeinstellung* R_i wird definiert als:

$$R_i = \sum_j \alpha_j \cdot F_{i,j}. \quad (9.5)$$

Der aggregierte Wert $F_{i,j}$ eines Merkmals j der Kameraeinstellung i wird mit den Faktoren α_j gewichtet, die individuelle Präferenzen eines Benutzers widerspiegeln. Beispielsweise ist es möglich, Gesichter besonders stark zu gewichten, so dass in einer Zusammenfassung nur Kameraeinstellungen mit großen Gesichtern gezeigt werden.

Die Auswahl der Kameraeinstellungen erfolgt als iterativer Prozess, der in Abbildung 9.4 ver-



Gesichter	0,38
Szenen	0,50
Bewegte Objekte	0,00
Kontrast	0,91
Bewegungsaktivität	0,20
Kamerabewegung	0,00
Zeitliche Verteilung	0,55
Gruppen ähnlicher	
Kameraeinstellungen	0,84
Summe	3,38

Gesichter	0,00
Szenen	0,50
Bewegte Objekte	0,00
Kontrast	0,94
Bewegungsaktivität	0,91
Kamerabewegung	0,00
Zeitliche Verteilung	1,00
Gruppen ähnlicher	
Kameraeinstellungen	0,69
Summe	4,04

Gesichter	0,00
Szenen	0,50
Bewegte Objekte	0,00
Kontrast	0,32
Bewegungsaktivität	0,09
Kamerabewegung	0,00
Zeitliche Verteilung	0,48
Gruppen ähnlicher	
Kameraeinstellungen	0,53
Summe	1,92

Abbildung 9.6: Beispiel für drei Kameraeinstellungen eines Zirkusfilms aus dem Jahre 1942. Die ersten beiden Kameraeinstellungen werden für die Zusammenfassung ausgewählt.

deutlich wird. Die aggregierten Merkmalswerte und der gewichtete Wert werden zunächst für alle Kameraeinstellungen berechnet. Die Kameraeinstellung mit dem maximalen Wert für R_i wird für die Zusammenfassung ausgewählt. Falls die Zusammenfassung noch nicht die gewünschte Länge erreicht hat, werden die dynamischen Merkmalswerte aktualisiert, und eine weitere Kameraeinstellung wird ausgewählt.

Abbildung 9.6 verdeutlicht am Beispiel von drei Kameraeinstellungen eines historischen Videos die aggregierten Merkmalswerte. Bei einer gleichmäßigen Gewichtung der Merkmalswerte werden die ersten beiden Kameraeinstellungen für die Zusammenfassung ausgewählt.

9.5.2 Überprüfung der ausgewählten Kameraeinstellungen

Einzelne Regeln müssen beachtet werden, damit eine qualitativ hochwertige Zusammenfassung erzeugt wird. Die ausgewählten Kameraeinstellungen werden dabei anhand folgender Regeln überprüft:

- Direkt aufeinander folgende Kamerabewegungen erzeugen einen unprofessionellen Eindruck des Videos, so dass Kameraeinstellungen mit deutlichen Kameraoperationen an Aufnahmen mit statischer Kamera angrenzen sollten.

- Zum besseren Verständnis der Handlung sollten mindestens zwei Kameraeinstellungen einer Szene ausgewählt werden.
- Die durchschnittliche Bewegungsaktivität sollte in der Zusammenfassung nicht wesentlich höher als im Originalvideo sein. Da die Zusammenfassung einer Video-Dokumentation einen vollständigen Überblick über das historische Video geben soll, ist eine zu starke Fokussierung auf schnelle Kameraeinstellungen nicht wünschenswert. Diese Regel ist bei Spielfilmen oder Sportveranstaltungen nicht anzuwenden, da Kameraeinstellungen mit hoher Bewegungsaktivität häufig besonders geeignet für diese Zusammenfassungen sind.
- Die Länge der Zusammenfassung sollte ungefähr der durch den Benutzer spezifizierten Länge entsprechen, wobei diese als absoluter oder relativer Wert festgelegt werden kann. Ohne Angabe der Länge wird sie innerhalb des *European-Chronicles-Online-Systems* mit zehn Prozent der Länge des ursprünglichen Videos festgelegt. Eine Anpassung erfolgt bei besonders kurzen oder langen historischen Videos, so dass die Länge einer Zusammenfassung immer zwischen einer und zehn Minuten liegt.
- Die Audiospur sollte nur in ruhigen Bereichen geschnitten werden.

Ist eine der Regeln verletzt, so werden in Abhängigkeit von der aktuellen Länge der Zusammenfassung einzelne Kameraeinstellungen entfernt, hinzugefügt oder ersetzt. Alle Bedingungen werden iterativ überprüft, bis keine Verletzung mehr auftritt bzw. bis die Summe der Fehlerwerte, welche die Verletzungen der einzelnen Regeln bewerten, nicht mehr abnimmt. Standardmäßig werden alle Fehler gleich gewichtet, wobei ein Anwender den einzelnen Bedingungen unterschiedliche Prioritäten zuweisen kann.

Falls individuelle Benutzerpräferenzen für die Erzeugung einer Zusammenfassung gewünscht sind, bleiben die Regeln unberücksichtigt. Erhalten beispielsweise Kameraeinstellungen mit starken Bewegungen eine besonders hohe Priorität, so wird die Bewegungsaktivität der Zusammenfassung deutlich über der des ursprünglichen Videos liegen, so dass mit hoher Wahrscheinlichkeit die entsprechende Regel verletzt wird.

Die Audiospur ist besonders wichtig für die Akzeptanz einer Zusammenfassung und wird nach der Überprüfung der Regeln gesondert betrachtet. Bei der Auswahl von zwei benachbarten Kameraeinstellungen bleibt die Audiospur unverändert. Im Falle eines Schnittes wird der am nächsten gelegene ruhige Bereich identifiziert und als Schnittposition ausgewählt. Liegt die Audio-Schnittposition weniger als fünf Sekunden von der durch die Bildinhalte ermittelte

Schnittposition entfernt, so werden einzelne Bilder der Kameraeinstellungen hinzugefügt bzw. entfernt, was einer Anpassung der Bildwiederholrate entspricht. Bei Kameraeinstellungen mit geringer Bewegungsaktivität und einer Länge von mehr als 30 Sekunden ist auch innerhalb dieser Kameraeinstellung an ruhigen Bereichen ein Schnitt zulässig. Wird keine geeignete Position zur Unterteilung der Audiospur gefunden, so wird die Audiospur innerhalb von 5 Sekunden ein- bzw. ausgeblendet. Durch die Überprüfung der ausgewählten Kameraeinstellungen und die Anpassung der Audiospur werden Zusammenfassungen der historischen Videos erzeugt, die wesentlich angenehmer zu betrachten sind.

9.5.3 Speicherung der Zusammenfassung

Im letzten Schritt werden die Übergänge zwischen den Kameraeinstellungen definiert, und die Zusammenfassung wird als Video¹ gespeichert. Dabei sollte der Anteil der Übergänge in der Zusammenfassung und dem ursprünglichen Video möglichst ähnlich verteilt sein. Physikalische Parameter des Videos, wie beispielsweise die Bitrate, die Bildauflösung oder die Bildwiederholrate, können durch den Anwender festgelegt werden. So kann beispielsweise aus einem Video in hoher Qualität im MPEG-II- oder MPEG-IV-Format eine Zusammenfassung als MPEG-I-Video mit deutlich geringerer Bitrate und Qualität erzeugt werden. Im Falle einer statischen Zusammenfassung werden die bei der Analyse des Videos ermittelten repräsentativen Einzelbilder der ausgewählten Kameraeinstellungen gespeichert.

9.6 Experimentelle Ergebnisse

Innerhalb der experimentellen Ergebnisse werden zunächst statische Zusammenfassungen betrachtet und beispielhaft sowohl eine in mehreren Systemen gewählte als auch eine neue Darstellungsform einer Zusammenfassung vorgestellt. Die Auswahl der Kameraeinstellungen ist für die statischen und dynamischen Zusammenfassungen identisch, lediglich das Abbruchkriterium wird durch die Anzahl und nicht durch die Länge der ausgewählten Kameraeinstellungen festgelegt. Die dynamischen Zusammenfassungen wurden in das *European-Chronicles-Online*-Projekt integriert und von professionellen Nutzern der Videoarchive evaluiert [276, 281].

¹Für die Ein- und Ausgabe werden die Formate MPEG-I, MPEG-II, MPEG-IV und Windows Media Video unterstützt.

9.6.1 Statische Zusammenfassungen von Videos

Bei einem naiven Ansatz zur Erstellung statischer Zusammenfassungen wird für jede Kameraeinstellung genau ein Bild ausgewählt und angezeigt. Der Nachteil liegt in der großen Anzahl Bilder, die viele Bildschirmseiten füllen bzw. so stark verkleinert werden müssen, dass der Inhalt nicht mehr erkannt werden kann. Durch Auswahl einzelner Kameraeinstellungen werden erst sinnvoll nutzbare Zusammenfassungen möglich. Abbildung 9.7 zeigt drei Beispiele für Zusammenfassungen von historischen Video-Dokumentationen.

Eine statische Zusammenfassung kann durch zusätzliche Informationen und spezielle Interaktionsmöglichkeiten erweitert werden. In vielen verfügbaren Systemen werden die dargestellten Bilder mit dem Video verknüpft, so dass der Betrachter durch Anklicken eines Bildes den entsprechenden Bereich des Videos betrachten kann. Um die Länge einer Kameraeinstellung und dessen Position innerhalb des Videos zu erkennen, hat es sich ebenfalls bewährt, diese als zusätzliche Information unter jedem Bild zu verdeutlichen (vgl. Abbildung 9.7).

Durch eine vergrößerte Darstellung einzelner Bilder können besonders wichtige Kameraeinstellungen hervorgehoben werden. Abhängig von der Anzahl der ausgewählten Bilder und dem verfügbaren Platz bietet sich die Darstellung in bis zu drei unterschiedlichen Skalierungsstufen an. In Abbildung 9.7 werden zwei Bildgrößen eingesetzt, wobei die Kameraeinstellung mit dem höchsten gewichteten Wert in voller Auflösung gezeigt wird und die übrigen Bilder auf 45 Prozent ihrer ursprünglichen Größe verkleinert werden.

Eine alternative und neue Darstellung einer statischen Zusammenfassung ist in Form einer *Kollage* möglich. Hierbei werden die einzelnen Kameraeinstellungen um einen Rahmen ergänzt und innerhalb eines größeren Bildes angeordnet, wobei es sich beim Hintergrund auch um ein Bild des Videos handelt. Abbildung 9.8 stellt zwei historische Videos als Kollage dar. In diesen Beispielen werden die verkleinerten Bilder gleichmäßig entlang zweier Bildränder angeordnet.

Das repräsentative Bild mit dem höchsten Kontrast wird als Hintergrundbild ausgewählt, da es in vielen der analysierten historischen Videos eine hohe Bildschärfe aufweist. Alternativ kann der Anwender ein Hintergrundbild aus der Liste aller repräsentativen Bilder auswählen. Auch das Layout, das durch die Anzahl und Anordnung der kleineren Bilder definiert ist, kann von einem Benutzer beeinflusst werden. Für die Berechnung der Abbildung 9.8 wurde lediglich das zu analysierende Video und die Anzahl der darzustellenden Bilder vorgegeben, die Auswahl, Anordnung und Berechnung der Kollage erfolgte automatisch.

Erweiterungen bei der Darstellung in Form einer Kollage sind möglich, indem beispielsweise



Abbildung 9.7: Beispiele einer statischen Zusammenfassung dreier historischer Videos aus den Jahren 1936 (a), 1937 (b) und 1939 (c). Die Positionen und Längen der ausgewählten Kameraeinstellungen werden durch einen Balken unter jedem Bild verdeutlicht.



Abbildung 9.8: Zwei statische Zusammenfassungen in Form von Kollagen

se Gesichter aus den Kameraeinstellungen automatisch ausgeschnitten und innerhalb eines größeren Bildes angeordnet werden. Bei der Zusammenfassung eines Spielfilms könnte zusätzlich durch Erkennung von Textregionen im Anfangsbereich des Filmes der Titel ermittelt und in die Kollage eingefügt werden. Bei einer Sportveranstaltung würden Nahaufnahmen der Sportler während besonderer Ereignisse wie beispielsweise eines Strafstoßes oder Torschusses angezeigt werden.

9.6.2 Dynamische Zusammenfassungen von Videos

Das Verfahren zur automatischen Erzeugung von dynamischen Zusammenfassungen wurde im Rahmen des Projektes *European Chronicles Online* von uns entwickelt und ist Teil des Systems zur Verwaltung und Indexierung historischer Videos. Die Sammlung der beteiligten Filmarchive enthält mehr als 100.000 Stunden historischer Videos, von denen zur Analyse der Zusammenfassungen mehr als 1.200 Videos aus den Jahren 1920 bis 1965 zur Verfügung stehen. Die Länge der Videos variiert zwischen einer und sechzig Minuten, wobei nur ab einer Länge von drei Minuten eine Zusammenfassung erzeugt wird.

Die Merkmale und aggregierten Merkmalswerte werden für jedes Video nur einmal berechnet und zur späteren Wiederverwendung als Metadaten im System gespeichert. Wird eine Zusammenfassung in anderer Länge oder mit veränderten Nutzerpräferenzen erzeugt, so können die Merkmale direkt aus der Datenbank des Systems ausgelesen werden. Die eigentliche Erzeugung und Speicherung der Zusammenfassung ist daher auf einem aktuellen PC fast in Echtzeit möglich, so dass der Anwender nach Spezifikation der neuen Präferenzen schon nach kurzer Zeit die entsprechende Zusammenfassung betrachten kann.

Neben der Evaluation mit professionellen Nutzern von Videoarchiven wurden erste Erfahrungen während der Entwicklung des Systems gesammelt. Zwei wesentliche Verbesserungsmöglichkeiten wurden in diesem Zusammenhang vorgeschlagen, die in das endgültige *European-Chronicles-Online*-System eingeflossen sind. Zum einen werden Kameraeinstellungen ohne sinnvollen Inhalt ausgewählt, in denen die Kamera beispielsweise zu Boden zeigt oder das Bild sehr unscharf ist. Da die meisten dieser Kameraeinstellungen einen sehr geringen Kontrast enthalten, wurde die Erkennung nicht relevanter Kameraeinstellungen entwickelt.

Die zweite Beobachtung betraf die Audiospur der Zusammenfassung, bei der eine Unterbrechung von Sprache oder Musik besonders unangenehm auffällt. Für historische Videos ist aufgrund des starken Rauschens innerhalb der Audiospur eine zuverlässige Spracherkennung mit heutiger Technik nicht möglich. Eine deutliche Verbesserung wird durch die Suche ruhiger

Bereiche bzw. das Ein- und Ausblenden des Audiosignals erreicht.

Insgesamt sind viele Anwender mit der Qualität der Zusammenfassungen sehr zufrieden. Trotz der deutlichen Verkürzung des Videos auf ungefähr zehn Prozent der ursprünglichen Länge bleiben in den meisten Zusammenfassungen wesentliche Teile des Inhaltes gut verständlich. Es wurde mehrfach beobachtet, dass bei sehr kurzen Zusammenfassungen wichtige Bestandteile des Videos ausgefiltert werden. Daher wurde die Mindestlänge für Zusammenfassungen auf eine Minute festgelegt.

Im Rahmen des Projektes *European Chronicles Online* wurde in einer zweitägigen Evaluation das System und die Qualität der automatisch erzeugten Zusammenfassungen analysiert. 17 professionelle Nutzer haben das System getestet, von denen fünf Personen im Bereich der Katalogisierung von Videos arbeiten und zwölf Personen für das Editieren der Videos zuständig sind. Ein wesentlicher Vorteil der Evaluation mit professionellen Nutzern liegt darin, dass sie die Aufgaben und Anforderungen an Archive sehr gut einschätzen können.

Während der zweitägigen Arbeit am System wurden Anmerkungen und Kommentare der Nutzer erfasst und durch Fragebögen und mündliche Interviews ergänzt. Allgemein wird die Qualität der Zusammenfassungen als sehr hoch eingeschätzt. Bei der Frage, ob die Arbeit mit den Archiven durch die automatisch erzeugten Zusammenfassungen unterstützt wird, schwanken die Ergebnisse jedoch deutlich (vgl. Abbildung 9.9). Innerhalb der Gruppe der Editoren haben mehrere Personen angemerkt, dass sie die Gefahr sehen, dass wesentliche Inhalte in der Zusammenfassung nicht berücksichtigt werden und der Inhalt verfälscht sein könnte. Fünf von zwölf Editoren haben geäußert, dass sie sich nicht auf automatisch erzeugte Zusammenfassungen verlassen wollen und die Arbeit mit dem Originalmaterial bevorzugen. Von den Katalogisierern werden automatisch generierte Zusammenfassungen als sehr positiv wahrgenommen. Im Interview äußerten mehrere Katalogisierer, dass sie sich vorstellen können, anhand der Zusammenfassungen kurze textuelle Beschreibungen des Videos zu erstellen, und dadurch eine deutliche Zeitersparnis bei ihrer Arbeit erwarten.

9.7 Zusammenfassung

In diesem Kapitel wurden neue Verfahren zur automatischen Erzeugung von Zusammenfassungen vorgestellt, die auf die besonderen Herausforderungen von historischen Dokumentationen eingehen. So verhindern die entwickelten Algorithmen, dass einzelne fehlerhafte Bilder als repräsentative Bilder einer Kameraeinstellung ausgewählt werden. Ein weiteres neues Verfahren zur Gruppierung von Kameraeinstellungen wurde entwickelt, bei dem fehlerhafte

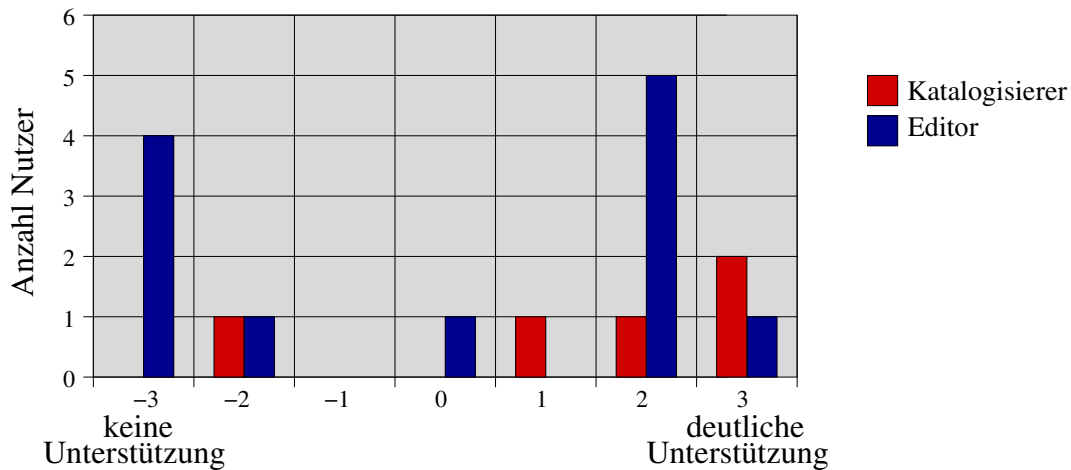


Abbildung 9.9: Antworten der Benutzer innerhalb der Evaluation auf die Frage: "Erwarten Sie, dass automatisch erzeugte Zusammenfassungen Ihre Arbeit unterstützen werden?"

Kameraeinstellungen defekten Gruppen zuordnet werden.

Zur Berechnung einzelner Merkmale wie beispielsweise der Bewertung von Kamerabewegungen, Szenen oder der zeitlichen Verteilung der ausgewählten Kameraeinstellungen wurden eine Vielzahl neuer Heuristiken vorgestellt. Anschließend erfolgte die Auswahl geeigneter Kameraeinstellungen, bei der zwei wesentliche neue Verfahren eingeführt wurden: die Auswahl nicht relevanter Kameraeinstellungen und die Kombination von festen und dynamisch veränderlichen Merkmalswerten. Im letzten Schritt des Algorithmus erfolgte eine Überprüfung aller ausgewählten Kameraeinstellungen anhand spezieller Regeln. Durch die Anwendung der Regeln konnten Zusammenfassungen von historischen Videos erzeugt werden, die deutlich angenehmer zu betrachten sind.

Im Rahmen der experimentellen Ergebnisse wurde am Beispiel einer Kollage eine neue Darstellungsform zur Präsentation statischer Zusammenfassungen vorgestellt. Anschließend wurde auf Evaluationsergebnisse eingegangen, wobei die Evaluation mit professionellen Nutzern von Videoarchiven durchgeführt wurde. Insbesondere mehrere Katalogisierer bewerteten die automatisch erzeugten Zusammenfassungen als sehr positiv und konnten sich vorstellen, diese zu nutzen, um textuelle Beschreibungen der Videos zu erstellen.

Abschließend lässt sich festhalten, dass die automatische Erzeugung von qualitativ hochwertigen Zusammenfassungen eine große Herausforderung darstellt. Obwohl objektive Kriterien – wie beispielsweise das Trennen der Audiospur innerhalb eines ruhigen Bereiches – berücksichtigt werden können, ist die Auswahl der Kameraeinstellungen sehr subjektiv. Eine optimale Zusammenfassung kann mit heutiger Technik nicht automatisch erzeugt werden, da *krea-*

tive und *künstlerische* Fähigkeiten ganz wesentlich bei der Erstellung eines Video einfließen. Selbst die Definition einer *optimalen Zusammenfassung* ist allgemein nicht möglich. Es ist zu erwarten, dass zwei Personen unterschiedliche Kameraeinstellungen eines längeren Videos auswählen und diese individuell kombinieren, da die einzelnen Kameraeinstellungen subjektiv unterschiedliche Bedeutungen haben. Eine automatisch erzeugte Zusammenfassung wird eine dritte Auswahl von Kameraeinstellungen treffen.

KAPITEL 10

Analyse der Bewegungen von Objekten und Personen

Im Gegensatz zur Objekterkennung in Videos, bei der die Frage im Mittelpunkt steht, welche Objekte im Bild dargestellt sind, werden mit der Bewegungsanalyse detaillierte semantische Informationen über ein Objekt ermittelt. Von besonderem Interesse sind Veränderungen eines Objektes im Zeitablauf, da aus diesen Daten wichtige semantische Informationen abgeleitet werden können. Hierzu zählen beispielsweise die Entfernung zur Kamera oder die Fahrtrichtung, die Geschwindigkeit und Richtungsänderungen eines PKWs.

In Videoarchiven werden häufig sehr spezielle Videosequenzen mit genau spezifizierten Inhalten gesucht. Ein Beispiel ist die Suche nach einem dunklen PKW, der innerhalb eines Zeitraumes von zehn Sekunden das Bild durchquert und sich dabei von der Kamera entfernt. Die in diesem Kapitel vorgestellten Algorithmen berechnen automatisch die zur Beantwortung dieser Suchanfrage benötigten Metadaten.

Neben der Analyse von Fahrzeugen sind in Videos insbesondere die Bewegungen und Gesten einer Person wichtig. Die traditionellen Interaktionsschemata zwischen Mensch und Maschine, die heute immer noch im Wesentlichen auf Tastatur und Maus beschränkt sind, könnten durch natürliche Interaktionsformen ersetzt werden. Bei der Kommunikation zwischen Menschen werden neben der Sprache wesentliche Informationen durch Gesten übermittelt, so dass die Auswertung dieser visuellen Informationen auch die Kommunikation mit einem Rechner verbessern würde.

Mehrere Anwendungen zur Analyse von Bewegungen einer Person sind verfügbar, bei denen Geräte durch Gesten gesteuert werden [190, 233, 237]. Insbesondere bei genau definierten Anwendungsgebieten lassen sich häufig die Fehlklassifikationen durch eine geringere Komplexität der Erkennungsalgorithmen reduzieren [148]. Einfache Gesten und Bewegungen des Kopfes wie beispielsweise Zustimmung oder Ablehnung können durch Analyse der Pupillen und der Positionsänderung des Kopfes zuverlässig erkannt werden [109]. Neben allgemein einsetzbaren Verfahren zur Erkennung von Gesten [299, 512, 544] liegt ein wichtiger Schwerpunkt in der Erkennung von Zeichensprache, die als spezielle Form der Gestenerkennung interpretiert werden kann [371, 468].

Eine Anwendung für Algorithmen zur Analyse der Bewegungen einer Person sind sogenannte *intelligente Räume* (engl. *smart room*), in denen Bewegungen und Gesten von Personen automatisch erkannt werden, um elektronische Geräte zu steuern [58, 415]. Ein Beispiel ist der *KidsRoom*, der computergesteuerte interaktive Spiele für Kinder ermöglicht [42, 43]. Das System analysiert die Videoströme von drei Kameras und wertet die Bewegungen der Kinder in Echtzeit aus. Gute Klassifikationsergebnisse sind möglich, da die Aktionen der Kinder durch die vorgegebene spielerische Handlung leicht vorhersehbar sind und der genaue Aufbau des Raumes und der enthaltenen Objekte bekannt ist.

Eine weitere Einsatzmöglichkeit für Algorithmen zur Analyse von Bewegungen und Gesten liegt im Bereich von *Überwachungssystemen* (engl. *surveillance*), die Personen nicht nur identifizieren oder mit mehreren Kameras verfolgen können, sondern auch spezielle Ereignisse und Aktivitäten automatisch erkennen. Insbesondere im öffentlichen Transportwesen wie Bahnhöfen oder Flughäfen, in Banken und Geschäften sowie in staatlichen Einrichtungen und Krankenhäusern setzt sich die Überwachung mit Videokameras zunehmend durch. Um das Sicherheitspersonal bei der Arbeit zu unterstützen, müssen die Algorithmen zur Analyse der Überwachungsvideos verdächtige Verhaltensweisen und Aktionen automatisch und in Echtzeit erkennen können [146]. Besonders wichtig ist die Identifikation von ungewöhnlichen Ereignissen und illegalen Aktivitäten wie beispielsweise einem Diebstahl oder Überfall [303, 588].

In diesem Kapitel wird ein neuer Ansatz zur Erkennung der Bewegungen von Objekten und Personen vorgestellt, der im Gegensatz zu den bisher dargelegten Verfahren auch für die Analyse von Videos geeignet ist. Im nächsten Abschnitt werden zunächst bekannte Verfahren zur Bewegungsanalyse erläutert, die für Videos jedoch nur eingeschränkt einsetzbar sind, da sie kalibrierte statische Kameras verwenden und eine nahezu fehlerfreie Segmentierung der Personen voraussetzen. Anschließend wird in Abschnitt 10.2 auf die besonderen Anforderungen bei der Analyse von Videos eingegangen und ein Überblick über das von uns entwickelte Ver-

fahren gegeben. Nach der Einführung der erweiterten Datenbank wird in Abschnitt 10.4 ein neuer Algorithmus zur Aggregation der Klassifikationsergebnisse vorgestellt, der insbesondere *Änderungen einer Kontur im Zeitablauf* berücksichtigt. In einer Übergangsmatrix wird ein Pfad mit minimalen Kosten gesucht, wobei die Wahrscheinlichkeiten der Übergänge zwischen Objektklassen berücksichtigt werden. Dadurch wird auch bei Fehlklassifikationen einzelner Bilder eine zuverlässige Erkennung der Bewegungsabläufe möglich. Abschließend werden experimentelle Ergebnisse in den Abschnitten 10.5 und 10.6 zur Analyse der Fahrt eines PKWs und der Bewegungen von Personen präsentiert.

10.1 Verfahren zur Analyse von Bewegungen

Frühere Ansätze zur Erkennung von Bewegungen einer Person haben Spezialhardware wie beispielsweise Handschuhe oder Sensoren an der Kleidung vorausgesetzt [300, 386, 478]. Mit zunehmender Rechenleistung und durch die Entwicklung neuer Algorithmen ist es heute möglich, zeitnah die Bilder einer oder mehrerer Kameras zu analysieren und so Bewegungen von Personen zu identifizieren. Eine Möglichkeit zur Erkennung der Gesten einer Person ist die Identifikation einzelner Körperteile wie beispielsweise Hände, Füße und Arme durch Analyse von Farb- und Konturinformationen [191, 232, 550]. Anhand der räumlichen Beziehungen der segmentierten Bildregionen werden die Positionen weiterer Körperteile abgeleitet [49]. Ausgehend von den unterschiedlichen Bewegungsrichtungen der einzelnen Körperteile können Aktivitäten von Personen abgeleitet werden [155, 156].

Sowohl zwei- als auch dreidimensionale Verfahren werden zur Analyse der Bewegungen einer Person eingesetzt. Im zweidimensionalen Fall kann die Konturanalyse als spezielle Form der Mustererkennung interpretiert werden, bei der Merkmale identifiziert und mit bekannten Mustern verglichen werden [164]. Durch den Einsatz mehrerer Kameras, die eine Person aus unterschiedlichen Richtungen aufnehmen, kann eine Person als dreidimensionales Modell erfasst werden [111, 222, 357]. Das Modell schränkt die zulässigen Bewegungen und Körperhaltungen ein und führt so zu einer Verbesserung der Klassifikationsergebnisse [389].

In mehreren Anwendungen werden Algorithmen zur Analyse der Bewegungen und Gesten einer Person eingesetzt. So ist es beispielsweise möglich, Gesten als zusätzliche Eingabemöglichkeit für Spiele zu verwenden [428], die Bewegung einer Person auf eine animierte Figur im Rechner in Echtzeit abzubilden [397] oder eine für einen Menschen natürlichere Kommunikation mit einem Roboter in Echtzeit zu ermöglichen [63]. Ein weiteres System bildet durch Projektion einen Touchscreen ab und ermöglicht mit Hilfe visueller Sensoren die Eingabe von

Daten. Durch Videoanalyse wird die exakte Position der Fingerspitzen im dreidimensionalen Raum bestimmt [344]. Auch im Bereich der Krankenpflege kann ein Videoüberwachungssystem Unterstützung bieten, indem zunächst Personengruppen wie beispielsweise Ärzte, Pfleger oder Patienten identifiziert und anschließend auffällige Aktivitäten für die einzelnen Gruppen erkannt werden [76].

10.2 Systemüberblick

Die vorgestellten Verfahren und Systeme zur Erkennung von Bewegungen sind für die Analyse von Videos nur bedingt geeignet. Viele Ansätze erfordern exakt segmentierte Objekte, die für Videosequenzen wegen der enthaltenen Kamerabewegung häufig nicht ausreichend genau zur Verfügung stehen. Bei Überwachungsvideos oder der Steuerung eines Rechners kann von einer statischen Kamera ausgegangen werden, so dass eine wesentlich genauere Segmentierung möglich ist und der Anteil der fehlerhaft klassifizierten Objekte stark sinkt. Mehrere Ansätze nutzen dreidimensionale Modelle des menschlichen Körpers. Für eine korrekte Abbildung einer segmentierten Person werden jedoch Tiefeninformationen benötigt, die aus einer einzelnen Kameraaufnahme nicht ermittelt werden können.

In Folgenden wird ein neues Verfahren zur Analyse der Bewegungen von Objekten und Personen in Videos vorgestellt, das nicht den oben genannten Einschränkungen unterliegt. Der wichtigste Schritt ist die Aggregation der Ergebnisse der einzelnen Bilder, durch die logische Zusammenhänge in Bewegungsabläufen abgebildet und fehlerhafte Klassifikationen ausgefiltert werden. Im Rahmen der Bewegungsanalyse wird die Drehung eines Objektes bzw. die Blickrichtung auf das Objekt, die Aktivität einer Person, die Entfernung zur Kamera sowie die Bewegungsrichtung und Geschwindigkeit des Objektes automatisch ermittelt.

Die Analyse der Bewegungen erfolgt entsprechend der vier in Abbildung 10.1 dargestellten Schritte. Zunächst werden innerhalb einer Kameraeinstellung sich bewegend Objekte mit dem in Kapitel 4 vorgestellten Verfahren segmentiert. Das zu segmentierende Objekt darf eine gewisse Größe nicht überschreiten, da bei sehr großen Objekten nicht zwischen Vordergrund und Hintergrund unterschieden werden kann und die Erzeugung des Hintergrundbildes fehlschlägt.

Die Erkennung des segmentierten Objektes erfolgt durch den Vergleich von Skalenraumabbildungen (vgl. Kapitel 5), bei denen zur Verbesserung der Klassifikationsergebnisse transformierte Konturen berücksichtigt werden. Um eine detailliertere Beschreibung eines Objektes zu erhalten, werden innerhalb der Datenbank die beiden Objektklassen *PKW* und *Person* in

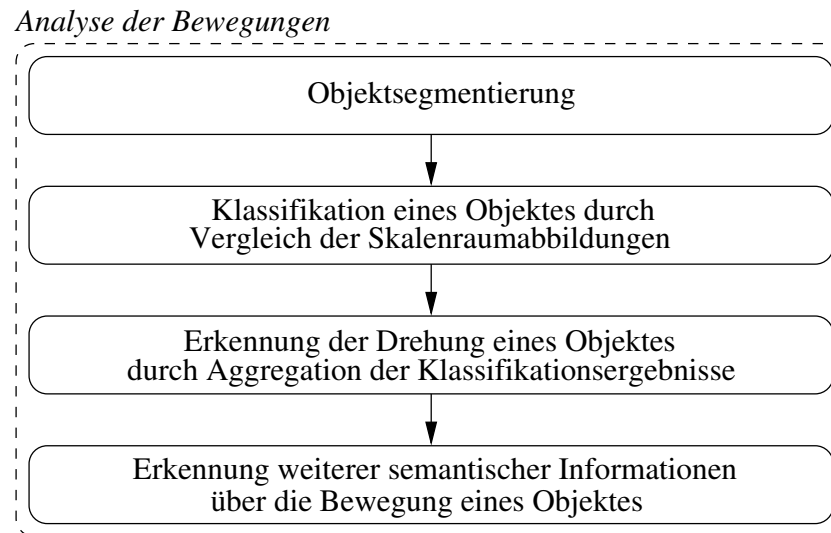


Abbildung 10.1: Analyse der Objekt- und Personenbewegungen

Unterklassen aufgeteilt.

Die Erkennung der Bewegung des Objektes innerhalb einer Kameraeinstellung erfolgt durch Aggregation der Klassifikationsergebnisse der einzelnen Bilder. Innerhalb einer Kameraeinstellung bewertet eine Kostenfunktion einen möglichen Wechsel eines Objektes zwischen den Unterklassen. Wahrscheinliche Änderungen verursachen niedrige Kosten wie beispielsweise der Wechsel einer Kontur von der Unterklasse *Person–gehen* in *Person–stehen*. Im letzten Schritt wird für das Objekt die Entfernung zur Kamera und die Geschwindigkeit der Bewegung ermittelt. Falls die Aufnahmeparameter, die physikalischen Eigenschaften der Kamera oder die Objektgrößen unbekannt sind, ist nur eine Annäherung der Entfernung bzw. der Geschwindigkeit durch geschätzte Parameter möglich.

10.3 Erweiterung der Datenbank

Die in Kapitel 5.9 eingeführte Datenbank mit den sechs Objektklassen *Säugetier*, *Vogel*, *Flugzeug*, *Schiff*, *PKW* und *Person* wird erweitert, um eine detailliertere Beschreibung eines Objektes zu ermöglichen. Zur Analyse der Bewegungen werden genauere Informationen benötigt, die in der Datenbank durch *Unterklassen* abgebildet werden. Objekte innerhalb einer Unterklasse beschreiben die Blickrichtung auf ein starres Objekt wie beispielsweise die frontale oder seitliche Aufnahme eines PKWs oder charakterisieren die Tätigkeit bzw. Bewegung einer Person. Tabelle 10.1 gibt einen Überblick über die Objektklassen und Unterklassen der

Name der Objektklasse	Anzahl der Elemente der Datenbank
Säugetier	38
Vogel	25
Flugzeug	22
Schiff	27
PKW	(63)
– frontal	12
– diagonal	36
– seitlich	15
Person	(137)
– Nachrichtensprecher	16
– gehen	64
– stehen	24
– sitzen	10
– drehen	11
– hinsetzen / aufstehen	12
Summe	312

Tabelle 10.1: Verteilung der Objekte der Datenbank auf die Objektklassen

Datenbank.

Die Objekte in einzelnen Unterklassen wie beispielsweise frontale PKWs oder Nachrichtensprecher variieren sehr wenig, so dass nur eine geringe Anzahl repräsentativer Objekte für diese Klassen benötigt wird. Besonders viele Objekte sind in der Unterklasse *Person–gehen* zusammengefasst, da sich diese Konturen durch die unterschiedlichen Positionen der Arme und Beine stark unterscheiden können. Obwohl nur Videosequenzen mit PKWs und Personen analysiert werden, bleiben die zusätzlichen Objektklassen in der Datenbank enthalten, um die Stabilität des Algorithmus zu überprüfen.

10.4 Aggregation der Klassifikationsergebnisse

Bei der Aggregation der Klassifikationsergebnisse werden insbesondere *Änderungen einer Kontur im Zeitablauf* durch Drehungen oder Verformungen des Objektes berücksichtigt [279]. Die Kontur eines Autos unterscheidet sich beispielsweise bei frontalen und seitlichen Aufnahmen deutlich. Noch stärkere Änderungen treten bei Konturen von Personen auf, da sowohl eine Drehung des Körpers als auch eine Änderung der Position der Arme und Beine möglich ist.

Zur Beschreibung der Übergänge zwischen Objektklassen werden Kosten definiert, die die

Wahrscheinlichkeiten für den Wechsel von einer Objektklasse bzw. Unterklasse in eine andere beschreiben. So wird ein seitlich sichtbarer PKW mit hoher Wahrscheinlichkeit auch im folgenden Bild von der Seite und mit deutlich geringerer Wahrscheinlichkeit aus der Diagonalen dargestellt sein. Die Wahrscheinlichkeit einer frontalen Aufnahme ist ohne vorherige diagonale Aufnahme äußerst gering und weist auf einen Segmentierungs- oder Klassifikationsfehler im aktuellen oder vorherigen Bild hin.

In einer Übergangsmatrix $w_{k,m}$ werden Kosten definiert, die ein Wechsel von Objektklasse k zu Objektklasse m verursacht, wobei es sich bei den Objektklassen auch um Unterklassen handeln kann. Durch den Vergleich der Skalenraumabbildungen sind die Differenzen zwischen den unbekannten Objekten der Kameraeinstellung und allen Objekten der Datenbank bekannt. Die minimale Differenz eines Objektes i wird für jede Objektklasse bzw. Unterklasse k in einer Matrix $d_{k,i}$ gespeichert.

Ziel ist es, die gesamten Kosten K zu minimieren, die sich aus den Kosten für die Übergänge $w_{k,m}$ zwischen zwei Objektklassen und den Kosten der Differenz $d_{k,i}$ eines Objektes zur Objektklasse zusammensetzen:

$$K = \min_c \sum_{i=1}^N d_{c_i,i} + w_{c_{i-1},c_i}. \quad (10.1)$$

Der Vektor c soll so bestimmt werden, dass die aggregierten Kosten für die Klassifikation aller Objekte und Übergänge innerhalb einer Kameraeinstellung minimal werden. Die Länge des Vektors c , der die erkannten Objektklassen für die einzelnen Bilder i beschreibt, entspricht der Anzahl der Bilder der Kameraeinstellung. Die Kosten der Übergänge $w_{k,m}$ werden so definiert, dass sie besonders hohe Werte zwischen unterschiedlichen Objektklassen annehmen und bei typischen Änderungen zwischen Unterklassen, wie beispielsweise dem Wechsel zwischen den Unterklassen *Person–stehen* und *Person–gehen*, nur geringe Differenzwerte erhalten.

Das Minimierungsproblem kann als Suche des kürzesten Pfades in einem Graphen interpretiert werden. Die Kosten $d_{k,i}$ repräsentieren die Knoten des Graphen, die Kanten entsprechen den Kosten $w_{k,m}$ für die Übergänge zwischen den Objektklassen. Abbildung 10.2 verdeutlicht die Struktur des Minimierungsproblems. Beginnend mit dem ersten Bild werden die minimalen Kosten der Übergänge zwischen den Objektklassen $w_{k,m}$ und der Differenz eines Objektes zur Objektklasse $d_{k,i}$ summiert. Beim Erreichen des letzten Bildes sind die minimalen Kosten für die gesamte Kameraeinstellung bekannt, und der Pfad kann in entgegengesetzter Richtung bis zum ersten Bild zurückverfolgt werden. Der Ansatz der dynamischen Programmierung wird

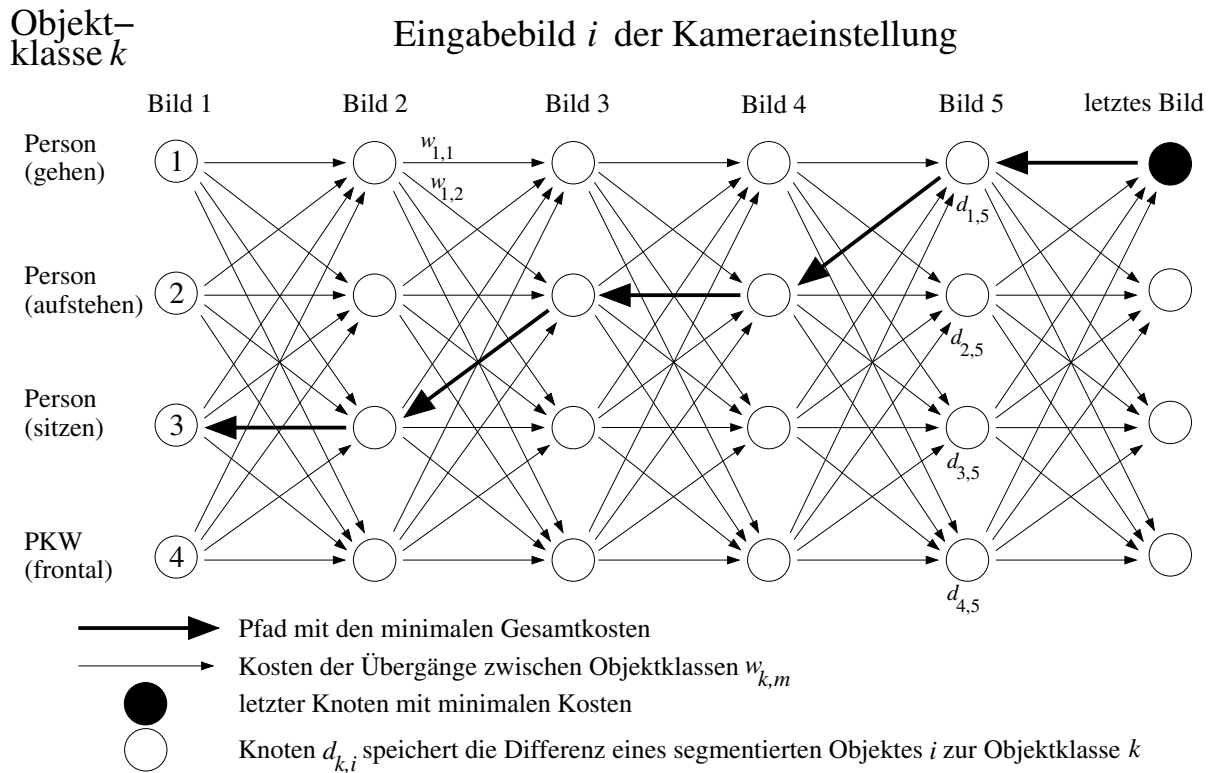


Abbildung 10.2: Ermittlung der Objektklasse mit Hilfe einer Übergangsmatrix





verwendet, um den optimalen Pfad im Graphen und somit die optimalen Übergänge zwischen den Objektklassen zu bestimmen [117, 136, 380].

10.5 Semantische Analyse der Fahrt eines PKWs





Bei der Analyse der Fahrt eines PKWs werden drei Blickrichtungen unterschieden, die durch Unterklassen mit *seitlichen*, *frontalen* oder *diagonalen* Aufnahmen von Fahrzeugen abgebildet werden. In der Übergangsmatrix sind geringe Kosten für den Wechsel zwischen den Unterklassen *PKW–seitlich* und *PKW–diagonal* bzw. *PKW–diagonal* und *PKW–frontal* festgelegt. Die Lösung des Minimierungsproblems gibt für jedes einzelne Bild der Kameraeinstellung den Objektnamen der Unterklasse an. Die Abbildung 10.3 verdeutlicht am Beispiel von drei Videosequenzen die Erkennung der Unterklassen für PKWs.






In Tabelle 10.2 werden die Ergebnisse mit und ohne Anwendung des Aggregationsalgorithmus verglichen. Der Anteil der fehlerhaft klassifizierten Einzelbilder in den drei Sequenzen mit den PKWs liegt bei 14 Prozent. Durch die Aggregation sinkt der Anteil auf unter drei Prozent. Die

Sequenz: PKW-1





				
Ansicht	diagonal	diagonal	diagonal	diagonal
Abstand	23,7 m	18,2 m	11,9 m	9,3 m
Geschwindigkeit	35 km/h	30 km/h	36 km/h	34 km/h

Sequenz: PKW-4

				
Ansicht	diagonal	diagonal	frontal	frontal
Abstand	5,9 m	7,1 m	7,5 m	9,0 m
Geschwindigkeit	21 km/h	15 km/h	14 km/h	17 km/h

					
Ansicht	diagonal	diagonal	diagonal	seitlich	seitlich
Abstand	12,0 m	12,8 m	14,8 m	15,5 m	15,6 m
Geschwindigkeit	24 km/h	31 km/h	29 km/h	36 km/h	43 km/h

Sequenz: PKW-5

				
Ansicht	diagonal	diagonal	diagonal	seitlich
Abstand	20,9 m	12,9 m	10,4 m	7,9 m
Geschwindigkeit	62 km/h	65 km/h	67 km/h	71 km/h





				
Ansicht	diagonal	diagonal	diagonal	diagonal
Abstand	7,8 m	9,1 m	12,7 m	15,8 m
Geschwindigkeit	70 km/h	65 km/h	67 km/h	63 km/h

Abbildung 10.3: Beispiele für die Analyse der Fahrt eines PKWs

Sequenz	Anzahl Bilder	Fehlerhaft klassifizierte Bilder ohne Aggregation	Fehlerhaft klassifizierte Bilder mit Aggregation
PKW-1	32	1 (3 %)	0 (0 %)
PKW-4	19	6 (32 %)	2 (11 %)
PKW-5	22	3 (14 %)	0 (0 %)
Summe / Durchschnitt	73	10 (14 %)	2 (3 %)
Person-4	29	5 (17 %)	1 (3 %)
Person-9	239	37 (15 %)	7 (3 %)
Person-14	35	14 (41 %)	6 (18 %)
Person-16	261	45 (17 %)	19 (7 %)
Summe / Durchschnitt	564	101 (18 %)	33 (6 %)

Tabelle 10.2: Experimentelle Ergebnisse zur Bewegungsanalyse

Fehler treten im Bereich der Übergänge zwischen diagonalen und frontalen Aufnahmen in der Sequenz *PKW-4* auf. Im Vergleich zur manuellen Klassifikation werden die beiden Übergänge zwischen den Unterklassen ein Bild zu früh bzw. ein Bild zu spät erkannt.

Neben der Drehung des Objektes zur Kamera können weitere Informationen automatisch ermittelt werden. Die *Farbe* des PKWs wird durch eine Histogrammanalyse bestimmt. Die Objektpixel aller Bilder sind durch den Segmentierungsschritt bekannt und werden in einem Histogramm zusammengefasst. Die dominante Farbe des Histogramms definiert die Farbe des Fahrzeugs. Die *Position* eines Fahrzeugs innerhalb des Bildes wird durch den Schwerpunkt der Objektpixel entsprechend der Gleichung 5.4 ermittelt. Da die Kamerabewegung und das Hintergrundbild der Kameraeinstellung aus dem Segmentierungsschritt bekannt sind, lässt sich die Richtung der Bewegung im Zeitablauf genau bestimmen. Durch Kombination der Positionsinformationen und der Objektklasse ist eine detaillierte Beschreibung der Bewegungen möglich.

Ohne Daten über die Objektgröße oder die physikalischen Merkmale der Kamera, wie beispielsweise der Brennweite, kann der *Abstand eines Fahrzeugs zur Kamera* nicht genau bestimmt werden. Das Verhältnis von Objektgröße zur Bildauflösung liefert jedoch eine Abschätzung der Entfernung. Die Größe S eines Objektes im Bild ist umgekehrt proportional zur Entfernung D :

$$D = \frac{F_C \cdot F_S}{S}. \quad (10.2)$$

Zur Beschreibung der Größe eines Objektes wird dessen Höhe verwendet, da sie bei PKWs und Personen deutlich weniger variiert als die Objektbreite. Der Skalierungsfaktor F_C ist abhängig

von der Brennweite und beschreibt die physikalischen Merkmale der Kamera, F_S spezifiziert die tatsächliche Höhe eines Objektes und wird bei der Analyse von PKWs mit 1,40 Meter geschätzt. Der Skalierungsfaktor F_C wurde experimentell mit Hilfe von Beispielaufnahmen ermittelt. Fehlerhafte Schätzungen bei den Skalierungsfaktoren führen zu einem entsprechenden relativen Fehler bei der Entfernung, wobei das Verhältnis der Änderungen der Entfernungen innerhalb einer Kameraeinstellung jedoch unbeeinflusst bleibt.

Da zu jedem Zeitpunkt die Bildposition und die Entfernung zur Kamera bekannt sind, kann die zurückgelegte Entfernung des Objektes zwischen zwei Bildern und somit auch die *Geschwindigkeit des Objektes* berechnet werden. Dabei wird die Annahme getroffen, dass sich das Objekt zwischen zwei benachbarten Bildern jeweils linear bewegt und keine vertikalen Bewegungen auftreten. Die zurückgelegte Entfernung U_i zwischen den Bildern $i - 1$ und i wird wie folgt angenähert:

$$U_i = \sqrt{[F_W \cdot (P_{x_i} - P_{x_{i-1}})]^2 + [D_i - D_{i-1}]^2}. \quad (10.3)$$

D_i definiert für das Bild i die Entfernung des Objektes zur Kamera, P_{x_i} die horizontale Pixelposition des Schwerpunktes des Objektes. Die horizontale Verschiebung wird mit dem Faktor F_W gewichtet, der aus der Bildgröße des Objektes und der tatsächlichen Objektgröße abgeleitet wird. Durch Multiplikation mit der Bildwiederholrate R des Videos wird die Geschwindigkeit des Objektes zum Zeitpunkt i mit folgender Formel angenähert:

$$V_i = R \cdot U_i. \quad (10.4)$$

Für die Beispielobjekte in Abbildung 10.3 sind die Blickrichtung auf das Fahrzeug, die Entfernung zur Kamera und die geschätzte Geschwindigkeit angegeben.

10.6 Semantische Analyse der Bewegung einer Person

Die Algorithmen zur Analyse von Videos mit PKWs können nach geringen Modifikationen auch zur Erkennung von Personen eingesetzt werden. Die Unterklassen beschreiben nicht nur die Blickrichtung der Kamera, sondern ermöglichen auch die Erkennung der Körperhaltung einer Person. Innerhalb der Übergangsmatrix sind besonders niedrige Kosten für den Wechsel zwischen den Unterklassen *Person–sitzen* und *Person–aufstehen*, *Person–aufstehen* und *Person–stehen* sowie *Person–gehen* und *Person–stehen* definiert.

Die Klassifikationsergebnisse mit und ohne Aggregation werden in Tabelle 10.2 gegenüberge-

stellt. Der Anteil der fehlerhaft klassifizierten Bilder liegt ohne Aggregation der Ergebnisse bei 18 Prozent und sinkt durch die Aggregation auf unter sechs Prozent. Gelegentlich unterscheiden sich die Helligkeitswerte der Kleidung der Person und des Hintergrundes nur minimal, so dass in mehreren benachbarten Bildern starke Segmentierungsfehler auftreten. Innerhalb der fehlerhaft klassifizierten Bilder kann der Zeitpunkt eines Übergangs von einer Objektklasse zur anderen nicht zuverlässig erkannt werden und führt zu Fehlern bei der Klassifikation. Abbildung 10.4 verdeutlicht für vier Testsequenzen die Ergebnisse der automatischen Klassifikation der Bewegungen von Personen.

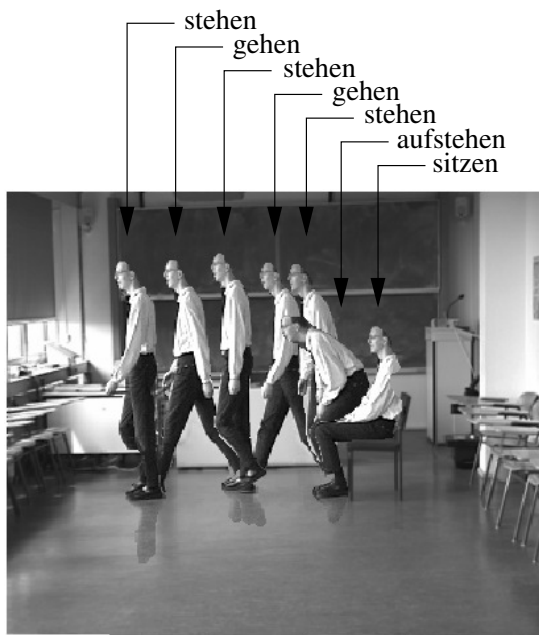
Bei der Identifikation der Kleidungsfarbe einer Person werden bei der Histogrammanalyse häufig zwei dominante Farben identifiziert, die den Farben der Hose und des Pullovers entsprechen. Zur Erhöhung der Genauigkeit der Klassifikation werden zwei getrennte Histogramme für die obere und die untere Objekthälfte erzeugt. Bei einfarbigen Kleidungsstücken entspricht die dominante Farbe eines Histogramms der Kleidungsfarbe, bei mehrfarbigen Kleidungsstücken ist keine zuverlässige Aussage möglich.

Am Beispiel der ersten beiden Testsequenzen in Abbildung 10.4, in denen die Entfernung zwischen Person und Kamera unverändert bleibt, wird deutlich, dass Schwankungen bei der Berechnung der Entfernung einer Person auftreten können. Die Fehler entstehen durch unterschiedliche Objektgrößen, da in einzelnen Bildern der Schatten im Bereich der Füße mit der Person segmentiert wird. Die Berechnung der Entfernung wird nur für die Unterklassen *Person–stehen*, *Person–gehen* und *Person–drehen* durchgeführt, da bei den anderen Objektklassen die Größenunterschiede zu stark variieren. Die Größe einer Person wird für diese drei Objektklassen mit 1,80 m angenommen.

10.7 Zusammenfassung

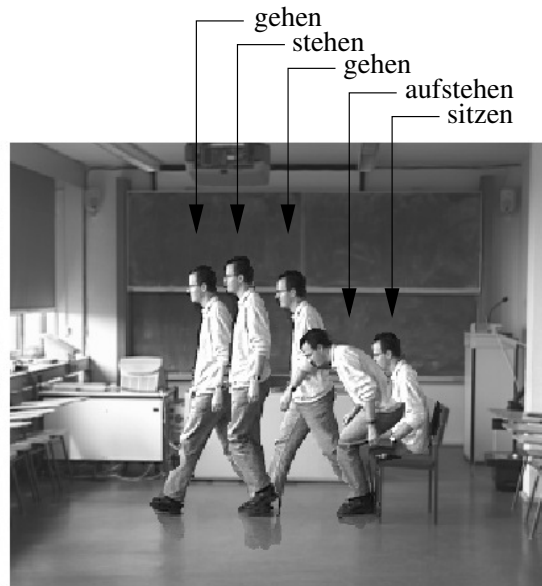
In diesem Kapitel wurde ein neuer Ansatz zur Erkennung der Bewegungen von PKWs und Personen vorgestellt, der im Gegensatz zu vielen bestehenden Verfahren nicht nur für Überwachungsszenarios, sondern auch zur Analyse von Videos geeignet ist. Das von uns entwickelte Verfahren ermöglicht es, detaillierte Beschreibungen der Bewegungen von Objekten und Personen in Videosequenzen automatisch zu ermitteln. Ein neuer Algorithmus zur Aggregation der Klassifikationsergebnisse wurde vorgestellt, der Veränderungen einer Kontur im Zeitablauf berücksichtigt. Dazu wurde eine Übergangsmatrix erstellt und der Pfad mit den minimalen Kosten berechnet, so dass trotz einer hohen Anzahl an fehlerhaft klassifizierten Einzelbildern eine zuverlässige und präzise Erkennung der Bewegungen möglich ist. Falls die Kameraparameter

Sequenz: Person-9



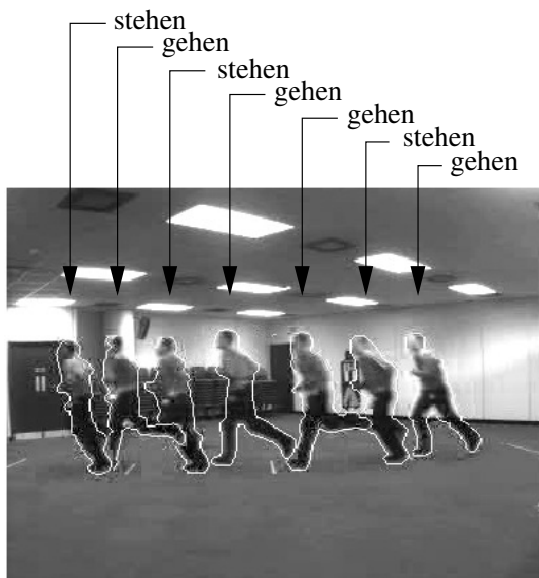
Abstand 4,7m 5,2m 4,7m 5,5m 5,4m unbekannt

Sequenz: Person-16



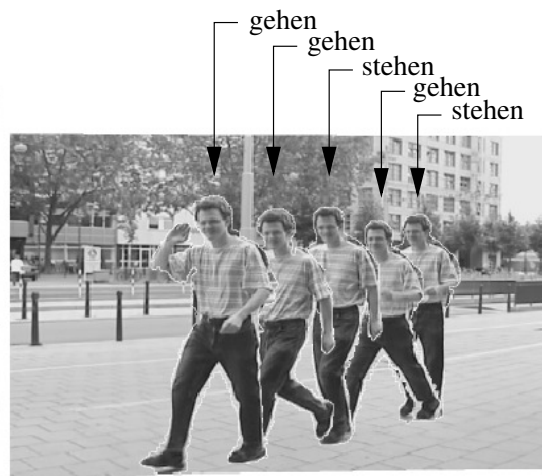
Abstand 5,2m 5,1m 5,3m unbekannt

Sequenz: Person-14



Abstand 9,9m 9,6m 9,7m 9,7m 9,5m 10,3m 10,6m

Sequenz: Person-4



Abstand 4,2m 4,5m 4,7m 5,3m 5,4m

Abbildung 10.4: Ergebnisse der Bewegungsanalyse von Personen

bekannt sind, kann die genaue Entfernung eines Objektes oder einer Person zur Kamera berechnet werden, ansonsten ist lediglich eine Schätzung möglich. Durch die Berechnung der Objektfarbe, der Position des Objektes im Bild und der Geschwindigkeit eines Objektes können weitere semantische Informationen über das Video automatisch ermittelt werden.

KAPITEL 11

Zusammenfassung und Ausblick

In dieser Arbeit wurden neue Algorithmen und Anwendungen zur Inhaltsanalyse von digitalen Videos vorgestellt. Die Analyseverfahren, die im ersten Teil der Arbeit erläutert wurden, bilden die Grundlage für die Anwendungen des zweiten Teils. Neben aktuellen Videos wurde zur Evaluation der entwickelten Algorithmen auf die umfangreiche Sammlung historischer Videos des Projektes *European Chronicles Online* zurückgegriffen. Mehrere Anwendungen und Analyseverfahren, wie beispielsweise die Erzeugung computergenerierter Zusammenfassungen oder die Schnitt-, Gesichts- und Objekterkennung, wurden in das *European-Chronicles-Online*-System integriert.

Im ersten Teil dieser Arbeit wurden zunächst Algorithmen zur *Schnitterkennung* erläutert. Neue Verfahren wurden entwickelt, um – trotz des großen Anteils an Bildfehlern – Schnitte zuverlässig in historischen Videos zu erkennen. Durch die Verbesserung können Werte von mehr als 90 Prozent für die Präzision und Vollständigkeit sowohl bei aktuellen als auch bei historischen Videos erreicht werden.

Bei der Analyse der *Kamerabewegung* wurde ein Verfahren erläutert, das eine sehr genaue Berechnung der Kameraparameter zwischen zwei benachbarten Bildern ermöglicht. Aus den Parametern des Kameramodells wurde eine allgemeine Beschreibung der Kamerabewegung abgeleitet, um Schwenks, Zoomoperationen oder eine verwackelte Kameraführung zu erkennen. Die Kameraparameter wurden in einem weiteren Schritt verwendet, um *Objekte zu segmentieren*, die sich vor dem Bildhintergrund bewegen. Durch die Ausrichtung aller Bilder einer Kameraeinstellung an einem Referenzbild wurde ein Hintergrundbild berechnet, in dem Objekte des Vordergrundes nicht mehr enthalten sind. Ein neues Verfahren zur Verringerung

der Fehler im Hintergrundbild wurde vorgeschlagen, bei dem die Position eines Objektes im Bild geschätzt wird und Objektpixel bei der Berechnung des Hintergrundbildes geringer gewichtet werden. Eine genaue Segmentierung wird durch morphologische Glättungsoperatoren und die Erkennung von Kanten im Randbereich des Objektes gewährleistet.

Im Rahmen der *Objekterkennung* wurden Skalenraumabbildungen zur Analyse der Kontur eines Objektes eingesetzt. Neue Algorithmen wurden in diesem für die Arbeit besonders wichtigen Kapitel präsentiert, die eine zuverlässige Klassifikation von Objekten ermöglichen. Mehrdeutigkeiten konkaver Objektregionen wurden vermieden, indem die Bogenbreite der Skalenraumabbildungen als zusätzliches Merkmal berücksichtigt wird. Zusätzlich sind durch die Einführung von transformierten Konturen Informationen über konvexe Objektregionen verfügbar. Die Algorithmen zur Segmentierung und Objekterkennung wurden in das *European-Chronicles-Online*-System integriert, so dass beim Einfügen eines neuen Videos Informationen über Objekte automatisch berechnet werden und den Anwendern des Archivs zur Verfügung stehen. Eine wesentliche Herausforderung bei der *Erkennung von Textregionen und Buchstaben* ist auf die geringe Bildauflösung eines Videos zurückzuführen. Ein neues Verfahren wurde entwickelt, das einen optimalen Pfad zwischen Buchstaben sucht und so geeignete Trenner zwischen den Buchstaben identifiziert. Eine deutliche Verbesserung der Segmentierung der einzelnen Buchstaben wird dadurch ermöglicht.

Gesichter sind von zentraler Bedeutung bei der computergestützten Analyse von digitalen Videos. Die *Gesichtserkennung* wurde als dreistufiges Verfahren implementiert, das aus der Lokalisierung einer Gesichtsregion, der Segmentierung und Normierung des Gesichtes sowie der eigentlichen Gesichtserkennung besteht. Bei den experimentellen Ergebnissen wurde auf semantische Fragestellungen, wie beispielsweise der gleichzeitig in einem Bild dargestellten Personen, eingegangen.

Im zweiten Teil der Arbeit wurden neue Anwendungen vorgestellt, welche die automatisch erkannten visuellen Inhalte eines Videos nutzen. Die erste Anwendung ermöglicht die *Adaption eines Videos*, bei der die Farbtiefe oder Bildauflösung angepasst wird. Zur Verringerung der Farbtiefe auf wenige Graustufenwerte wurde die Helligkeitsverteilung der Pixel einer Kameraeinstellung berücksichtigt. Für binäre Displays, die nur zwei unterschiedliche Helligkeitswerte anzeigen können, wurden durch die Überlagerung von Texturen und Kantenbildern besonders gute Ergebnisse erzielt. Die Anpassung der Bildauflösung erfolgte durch Bewertung der semantischen Inhalte einer Kameraeinstellung, aus denen der Bildausschnitt des adaptierten Videos abgeleitet wird. Speziell für historische Videos wurden Verfahren zur Verbesserung der Bildqualität vorgestellt, um die Helligkeit und den Kontrast anzupassen, Streifen und Kratzer

zu entfernen oder verwackelte Aufnahmen zu stabilisieren.

In einer zweiten Anwendung wurden neue Algorithmen zur automatischen Erzeugung von *Zusammenfassungen eines Videos* vorgestellt. Die Darstellung der Zusammenfassung ist als Liste mit einzelnen Bildern, als Kollage oder als Videosequenz möglich. Die Auswahl der Bilder oder Kameraeinstellungen hängt von den semantischen Inhalten des Videos ab. Eine Evaluation der automatisch erzeugten Zusammenfassungen mit sehr positiven Rückmeldungen der professionellen Anwender der Archive wurde im Rahmen des *European-Chronicles-Online*-Projektes durchgeführt.

Im letzten Kapitel wurde eine Anwendung entwickelt, um *Bewegungen von Objekten oder Personen* zu analysieren. Mögliche Veränderungen eines Objektes zwischen benachbarten Bildern wurden mit Hilfe einer Übergangsmatrix abgebildet. Durch die Analyse der Bewegungen im Zeitablauf werden detaillierte Informationen über die Geschwindigkeit, Bewegungsrichtung und die Art der Bewegung eines Objektes ermittelt.

Durch die *computergestützte Inhaltsanalyse von digitalen Videoarchiven* können wichtige semantische Informationen in Videos automatisch erkannt werden. Die Informationen erleichtern die Arbeit der Archivare und verbessern die Suchmöglichkeiten in den Archiven. Eine effiziente Suche nach Videos gewinnt auch außerhalb der Videoarchive zunehmend an Bedeutung. Ein Beispiel für eine Anwendung, die sich zur Zeit in der Entwicklung befindet, ist *Video Google* [540]. Die Anwendung soll die Veröffentlichung von Videos über das Internet unterstützen, digitales Rechtemanagement beinhalten und eine Komponente zur Abrechnung und Bezahlung von Videos zur Verfügung stellen. Die zentrale Funktionalität von *Video Google* ist jedoch die textbasierte Suchfunktion, die aktuelle Filme oder Serien von Fernsehsendern und Amateurvideos aus dem Internet mit Hilfe von Metadaten findet.

Es ist zu erwarten, dass die Bedeutung digitaler Videos in den nächsten Jahren weiter zunehmen wird und immer mehr Inhalte der Fernsehsender über das Internet abrufbar sind. Dadurch wird neben der Suche von Inhalten auch die Art der Darstellung eines Videos an Bedeutung gewinnen. Algorithmen zur automatischen Adaption und computergenerierte Zusammenfassungen sind erste Beispiele für neue Anwendungen in diesem Umfeld.

Referenzen

- [1] ABBASI, S. und F. MOKHTARIAN: *Shape Similarity Retrieval under Affine Transform: Application to Multi-View Object Representation and Recognition*. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, S. 450–455. IEEE Computer Society Press, 1999.
- [2] ABBASI, S., F. MOKHTARIAN und J. KITTLER: *Enhancing CSS-based shape retrieval for objects with shallow concavities*. In: *Image and Vision Computing*, Bd. 18(3), S. 199–211, 2000.
- [3] AGGARWAL, J. und N. NANDHAKUMAR: *On the computation of motion from sequences of images – A review*. In: *Proceedings of the IEEE*, Bd. 76(8), S. 917–935. IEEE Computer Society Press, August 1988.
- [4] AGUI, T., Y. KOKUBO, H. NAGASHASHI und T. NAGAO: *Extraction of face recognition from monochromatic photographs using neural networks*. In: *Proceedings of International Conference on Automation, Robotics and Computer Vision*, Bd. 1, S. 1881–1885, 1992.
- [5] ALATAN, A. A., A. N. AKANSU und W. WOLF: *Multi-Modal Dialog Scene Detection Using Hidden Markov Models for Content-Based Multimedia Indexing*. In: *Multimedia Tools and Applications*, Bd. 14(2), S. 137–151. Kluwer Academic Publishers, Juni 2001.
- [6] ALDINGER, T., S. KOPF, N. SCHEELE und W. EFFELSBERG: *Participatory Simulation of a Stock Exchange*. In: *World Conference on Educational Multimedia, Hypermedia and Telecommunications (EdMedia)*, S. 1–8, Montréal, Canada, September 2005.
- [7] AMER, A., E. DUBOIS und A. MITICHE: *Rule-based real-time detection of context-independent events in video shots*. In: *Elsevier Journal for Real-Time Imaging*, Bd. 11(3), S. 244–256, 2005.
- [8] AMIR, A., D. PONCELEON, B. BLANCHARD, D. PETKOVIC, S. SRINIVASAN und G. COHEN: *Using Audio Time Scale Modification for Video Browsing*. In: *IEEE Hawaii International Conference on System Sciences*, Bd. 3, S. 3046–3055. IEEE Computer Society Press, 2000.
- [9] ANDERSON, J. A., J. W. SILVERSTEIN, S. A. RITZ und R. S. JONES: *Distinctive features, categorical perception, and probability learning: some applications of a neural model*. In: *Neurocomputing*, S. 283–325. MIT Press, 1988.
- [10] ANER, A. und J. R. KENDER: *Video Summaries through Mosaic-Based Shot and Scene Clustering*. In: *Proceedings of the 7th European Conference on Computer Vision – Part IV*, Bd. 2353, S. 388–402, 2002.

- [11] ANER, A., L. TANG und J. R. KENDER: *A Method and Browser for Cross-Referenced Video Summaries*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Bd. 2, S. 237–240. IEEE Computer Society Press, 2002.
- [12] ANER-WOLF, A. und J. R. KENDER: *Video summaries and cross-referencing through mosaic-based representation*. In: *Computer Vision and Image Understanding*, Bd. 95(2), S. 201–237. Elsevier Science Inc., August 2004.
- [13] ANTANI, S., D. CRANDALL und R. KASTURI: *Robust Extraction of Text in Video*. In: *Proceedings of International Conference on Pattern Recognition (ICPR)*, S. 831–834, September 2000.
- [14] ANTANI, S., D. CRANDALL, A. NARASIMHAMURTHY, V. MARIANO und R. KASTURI: *Evaluation of Methods for Detection and Localization of Text in Video*. In: *Preproceedings of the IAPR Workshop on Document Analysis Systems*, S. 507–514, Dezember 2000.
- [15] ANTANI, S., R. KASTURI und R. JAIN: *A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video*. In: *Pattern Recognition*, Bd. 35(4), S. 945–965, 2002.
- [16] ARANDJELOVIC, O. und A. ZISSERMAN: *Automatic Face Recognition for Film Character Retrieval in Feature-Length Films*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 860–867. IEEE Computer Society Press, 2005.
- [17] ARMAN, F. und J. K. AGGARWAL: *Model-based object recognition in dense-range images—a review*. In: *ACM Computing Surveys (CSUR)*, Bd. 25 (1), S. 5–43. ACM Press, März 1993.
- [18] AUGUSTEIJN, M. und T. SKUJCA: *Identification of Human Faces through Texture-Based Feature Recognition and Neural Network Technology*. In: *Proceedings of IEEE Conference on Neural Networks*, S. 392–398. IEEE Computer Society Press, 1993.
- [19] BAASE, S. und A. V. GELDER: *Computer Algorithms: Introduction to Design and Analysis*. Addison-Wesley, Harlow, Essex, England, 3. Aufl., 1999.
- [20] BAI, B. und J. HARMS: *A multiview video transcoder*. In: *Proceedings of the 13th annual ACM international conference on Multimedia*, S. 503–506. ACM Press, 2005.
- [21] BAIRD, L.: *Reinforcement Learning Through Gradient Descent*. Techn. Ber. CMU-CS-99-132, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA, Mai 1999.
- [22] BALLARD, D. und C. BROWN: *Computer Vision*. Prentice-Hall, New Jersey, 1982.
- [23] BANHAM, M. R. und A. K. KATSAGGELOS: *Digital image restoration*. In: *IEEE Signal Processing Magazine*, Bd. 14 (2), S. 24–41. IEEE Computer Society Press, März 1997.
- [24] BARRETT, W. A.: *A survey of face recognition algorithms and testing results*. In: *Systems and Computers*, Bd. 1, S. 301–305, 1998.
- [25] BARRON, J. L., D. J. FLEET und S. S. BEAUCHEMIN: *Performance of Optical Flow Techniques*. In: *International Journal on Computer Vision*, Bd. 12(1), S. 43–77, 1994.

- [26] BATTIATO, S., D. CANTONE, D. CATALANO, G. CINCOTTI und M. HOFRI: *An Efficient Algorithm for the Approximate Median Selection Problem*. In: *Proceedings of Italian Conference on Algorithms and Complexity (CIAC)*, S. 226–238, März 2000.
- [27] BEAUCHEMIN, S. S. und J. L. BARRON: *The Computation of Optical Flow*. In: *ACM Computing Surveys*, Bd. 27(3), S. 433–467. ACM Press, 1995.
- [28] BEEK, P., J. R. SMITH, T. EBRAHIMI, T. SUZUKI und J. ASKELOF: *Metadata-driven multimedia access*. In: *IEEE Signal Processing Magazine*, Bd. 20(2), S. 40–52. IEEE Computer Society Press, März 2003.
- [29] BELHUMEUR, P., J. HESPAÑA und D. KRIEGMAN: *Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 19(7), S. 711–720. IEEE Computer Society Press, Juli 1997.
- [30] BELONGIE, S., M. J. und J. PUZICHA: *Matching shapes*. In: *IEEE International Conference on Computer Vision (ICCV)*, Bd. 1, S. 454–461. IEEE Computer Society Press, 2001.
- [31] BELONGIE, S., J. MALIK und J. PUZICHA: *Shape matching and object recognition using shape contexts*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 24, S. 509–522. IEEE Computer Society Press, April 2002.
- [32] BERTINI, M., R. CUCCHIARA, A. BIMBO und A. PRATI: *An Integrated Framework for Semantic Annotation and Adaptation*. In: *Multimedia Tools and Applications*, Bd. 26(3), S. 345–363. Springer Science & Business Media B.V., August 2005.
- [33] BERTSEKAS, D. P. und J. N. TSITSIKLIS: *Gradient Convergence In Gradient Methods With Errors*. Techn. Ber. LIDS-P-2404, Dept. of Electrical Engineering and Computer Science, MIT, Cambridge, MA, USA, 1997.
- [34] BICHSEL, M.: *Human Face Recognition: From Views to Models - From Models to Views*. In: *Proceedings of International Workshop on Automatic Face- and Gesture-Recognition (IWAAGR)*, S. 59–64, 1995.
- [35] BICHSEL, M.: *Automatic Interpolation and Recognition of Face Images by Morphing*. In: *Proceedings of International Conference on Automatic Face and Gesture Recognition (ICAFGR)*, S. 128–135, 1996.
- [36] BIEDERMAN, I.: *Recognition-by-components: a theory of human image understanding*. In: *Psychological Review*, Bd. 94, S. 115–147, 1987.
- [37] BIGÜN, J., G. H. GRANLUND und J. WIKLUND: *Multidimensional orientation estimation with applications to texture analysis and optical flow*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 13, S. 775–790. IEEE Computer Society Press, 1991.
- [38] BJÖRK, N. und C. CHRISTOPOULOS: *Video transcoding for universal multimedia access*. In: *Proceedings of the 2000 ACM workshops on Multimedia*, S. 75–79. ACM Press, 2000.
- [39] BLANZ, V. und S. ROMDHANI: *Face Identification across Different Poses and Illuminations with a 3D Morphable Model*. In: *Proceedings of the 5th IEEE International Conference on Automatic Face and Gesture Recognition*, S. 202–207. IEEE Computer Society Press, 2002.

- [40] BLANZ, V., M. TARR, H. BÜLTHOFF und T. VETTER: *What object attributes determine canonical views?*. Techn. Ber. No. 42, Max-Planck-Institut für Biological Cybernetics, Tübingen, Germany, 1996.
- [41] BOBER, M.: *MPEG-7 visual shape descriptors*. In: *IEEE Transactions on Circuits and Systems for Video Technology*, Bd. 11(6), S. 716–719. IEEE Computer Society Press, 2001.
- [42] BOBICK, A. F., S. S. INTILLE, J. W. DAVIS, F. BAIRD, C. S. PINHANEZ, L. W. CAMPBELL, Y. A. IVANOV, A. SCHÜTTE und A. WILSON: *The KidsRoom: A perceptually-based interactive and immersive story environment*. In: *PRESENCE: Teleoperators and Virtual Environments*, Bd. 8(4), S. 367–391, August 1999.
- [43] BOBICK, A. F., S. S. INTILLE, J. W. DAVIS, F. BAIRD, C. S. PINHANEZ, L. W. CAMPBELL, Y. A. IVANOV, A. SCHÜTTE und A. WILSON: *Perceptual user interfaces: the KidsRoom*. In: *Communications of the ACM*, Bd. 43 (3), S. 6–61. ACM Press, März 2000.
- [44] BOCCIGNONE, G., A. CHIANESE, V. MOSCATO und A. PICARIELLO: *Foveated Shot Detection for Video Segmentation*. Techn. Ber. 2, University of Salerno, Baronissi, Italy, 2005.
- [45] BOISSONNAT, J.-D. und M. YVINEC: *Algorithmic Geometry*. Cambridge University Press, Cambridge, New York, Melbourne, 1998.
- [46] BOKSER, M.: *Omnidocument Technologies*. In: *Proceedings of the IEEE*, Bd. 80(7), S. 1066–1078. IEEE Computer Society Press, Juli 1992.
- [47] BOLT, B. und D. HOBBS: *A Mathematical Dictionary for Schools*. Cambridge University Press, Cambridge, England, 1998.
- [48] BORECZKY, J., A. GIRGENSOHN, G. GOLOVCHINSKY und S. UCHIHASHI: *An Interactive Comic Book Presentation for Exploring Video*. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, S. 185–192. ACM Press, 2000.
- [49] BOULAY, B., F. BREMOND und M. THONNAT: *Human Posture Recognition in Video Sequence*. In: *Proceedings of Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, S. 23–29. IEEE Computer Society Press, Oktober 2003.
- [50] BOWYER, K. W., K. CHANG und P. J. FLYNN: *A survey of 3D and multi-modal 3D+2D face recognition*. In: *International Conference on Pattern Recognition (ICPR)*, S. 358–361, August 2004.
- [51] BRETSCHNEIDER, T., O. KAO und P. J. BONES: *Removal of Vertical Scratches in Digitised Historical Film Sequences Using Wavelet Decomposition*. In: *Proceedings of Image and Vision Computing*, S. 38–43, 2000.
- [52] BRETSCHNEIDER, T., C. MILLER und O. KAO: *Interpolation of scratches in motion picture films*. In: *International Conference on Acoustics, Speech, and Signal Processing*, Bd. 3, S. 1873–1876, 2001.

- [53] BROCKETT, R. W. und P. MARAGOS: *Evolution Equations for Continuous-Scale Morphological Filtering*. In: *IEEE Transactions Signal Processing*, Bd. 42(12), S. 3377–3386. IEEE Computer Society Press, Dezember 1994.
- [54] BRONSTEIN, A. M., M. M. BRONSTEIN und R. KIMMEL: *Three-Dimensional Face Recognition*. In: *International Journal of Computer Vision (IJCV)*, Bd. 64(1), S. 5–30. Springer Verlag, August 2005.
- [55] BROWN, L. G.: *A Survey of Image Registration Techniques*. In: *ACM Computing Surveys*, Bd. 24(4), S. 325–376. ACM Press, Dezember 1992.
- [56] BROWNE, P. und A. F. SMEATON: *Video information retrieval using objects and ostensive relevance feedback*. In: *Proceedings of the 2004 ACM symposium on Applied computing*, S. 1084–1090. ACM Press, 2004.
- [57] BROWNE, P., A. F. SMEATON, N. MURPHY, N. O’CONNOR, S. MARLOW und C. BERRUT: *Evaluation and combining digital video shot boundary detection algorithms*. In: *Proceedings of Irish Machine Vision and Information Processing Conference*, S. 93–100, 2000.
- [58] BRUMITT, B., B. MEYERS, J. KRUMM, A. KERN und S. SHAFER: *EasyLiving: Technologies for Intelligent Environments*. In: *Proceedings of the 2nd international symposium on Handheld and Ubiquitous Computing*, Bd. 1927, S. 12–29. Springer-Verlag, September 2000.
- [59] BRUNELLI, R. und D. FALAVIGNA: *Person Identification Using Multiple Cues*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 17(10), S. 955–966. IEEE Computer Society Press, Oktober 1995.
- [60] BRUNELLI, R. und T. POGGIO: *HyberBF Networks for Real Object Recognition*. In: *International Joint Conference on Artificial Intelligence*, S. 311–314, 1991.
- [61] BUHMANN, J., M. LADES und C. VON DER MALSBURG: *Size and distortion invariant object recognition by hierarchical graph matching*. In: *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, Bd. 2, S. 411–416, 1990.
- [62] BURL, M., T. LEUNG und P. PERONA: *Face Localization via Shape Statistics*. In: *Proceedings of International Workshop on Automatic Face and Gesture Recognition*, S. 154–159, Juni 1995.
- [63] BÖHME, H.-J., U.-D. BRAUMANN, A. CORRADINI und H.-M. GROSS: *Person Localization and Posture Recognition for Human-Robot Interaction*. In: *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, Bd. 1739, S. 117–128. Springer-Verlag, 1999.
- [64] BÜLTHOFF, H., S. EDELMAN und M. TARR: *How are three-dimensional objects represented in the brain?*. Techn. Ber. CogSci Memo No. 5, Max-Planck-Institut for Biological Cybernetics, Tübingen, Germany, 1994.
- [65] CABEDO, X. U. und S. K. BHATTACHARJEE: *Shot detection tools in digital video*. In: *Proceedings of Non-linear Model Based Image Analysis*, S. 121–126. Springer Verlag, Juli 1998.

- [66] CAI, M., J. SONG und M. LYU: *A New Approach for Video Text Detection*. In: *IEEE International Conference On Image Processing*, S. 117–120. IEEE Computer Society Press, September 2002.
- [67] CALIC, J. und E. IZQUIERDO: *Efficient Key-Frame Extraction and Video Analysis*. In: *International Conference on Information Technology: Coding and Computing*, S. 28–33, 2002.
- [68] CAMPISI, P., A. NERI und L. SORGI: *Automatic dissolve and fade detection for video sequences*. In: *International Conference on Digital Signal Processing (DSP)*, Bd. 2, S. 567–570, Juli 2002.
- [69] CAMPISI, P., A. NERI und S. SORGI: *Wipe effect detection for video sequences*. In: *Proceedings of IEEE 2002 Workshop on Multimedia Signal Processing (MMSP2002)*, S. 161–164. IEEE Computer Society Press, Dezember 2002.
- [70] CANNY, J. F.: *Finding Edges and Lines in Images*. Diplomarbeit, Massachusetts Institute of Technology, Juni 1983.
- [71] CANNY, J. F.: *A Computational Approach to Edge Detection*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 8(6), S. 679–698. IEEE Computer Society Press, 1986.
- [72] CARDELLINI, V., P. YU und Y. HUANG: *Collaborative Proxy System for Distributed Web Content Transcoding*. In: *Proceedings of 9th International ACM Conference on Information and Knowledge Management*, S. 520–527. ACM Press, November 2000.
- [73] CASTLEMAN, K. R.: *Digital Image Processing*. Prentice-Hall, New Jersey, 1996.
- [74] CERNEKOVA, Z., C. NIKOU und I. PITAS: *Entropy Metrics used for Video Summarization*. In: *International Spring Conference on Computer Graphics*, S. 1–8, April 2002.
- [75] CHELLAPPA, R., C. WILSON und S. SIROHEY: *Human and Machine Recognition of Faces: A Survey*. In: *Proceeding of the IEEE*, Bd. 83(5), S. 704–740. IEEE Computer Society Press, 1995.
- [76] CHEN, D., R. MALKIN und J. YANG: *Multimodal detection of human interaction events in a nursing home environment*. In: *Proceedings of the 6th international conference on Multimodal interfaces (ICMI)*, S. 82–89. ACM Press, 2004.
- [77] CHEN, H.-W., J.-H. KUO, W.-T. CHU und J.-L. WU: *Action movies segmentation and summarization based on tempo analysis*. In: *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, S. 251–258. ACM Press, 2004.
- [78] CHEN, L.-Q., X. XIE, X. FAN, W.-Y. MA, H.-J. ZHANG und H.-Q. ZHOU: *A visual attention model for adapting images on small displays*. In: *ACM Multimedia Systems Journal*, Bd. 9(4), S. 353–364. ACM Press, 2003.
- [79] CHEN, S.: *Quicktime VR – An image based approach to virtual environment navigation*. In: *Proceedings of Computer graphics and interactive techniques*, S. 29–38. ACM Press, 1995.
- [80] CHEONG, L. F. und H. GUO: *Shot Change Detection Using Scene-based Constraint*. In: *Multimedia Tools and Applications*, Bd. 14 (2), S. 175–186. Kluwer Academic Publishers, Juni 2001.

- [81] CHETVERIKOV, D. und A. LERCH: *Multiresolution Face Detection*. In: *Theoretical Foundations of Computer Vision*, Bd. 69, S. 131–140, 1993.
- [82] CHIMITT, W. J. und L. G. HASSEBROOK: *Scene reconstruction from partially overlapping images with use of composite filters*. In: *Journal of Optical Society of America A (JOSA)*, Bd. 16(9), S. 2124–2135, September 1999.
- [83] CHRISTEL, M. G.: *Visual digests for news video libraries*. In: *Proceedings of the 7th ACM international conference on Multimedia*, S. 303–311. ACM Press, 1999.
- [84] CHRISTEL, M. G., A. G. HAUPTMANN, H. D. WACTLAR und T. D. NG: *Collages as dynamic summaries for news video*. In: *Proceedings of the 2002 ACM workshops on Multimedia*, S. 561–569. ACM Press, 2002.
- [85] CHRISTEL, M. G., A. G. HAUPTMANN, A. S. WARMACK und S. A. CROSBY: *Adjustable Filmstrips and Skims as Abstractions for a Digital Video Library*. In: *Proceedings of the IEEE Advances in Digital Libraries Conference*, S. 98–104. IEEE Computer Society Press, 1999.
- [86] CHRISTEL, M. G., M. A. SMITH, C. R. TAYLOR und D. B. WINKLER: *Evolving video skims into useful multimedia abstractions*. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, S. 171–178. ACM Press/Addison-Wesley Publishing Co., April 1998.
- [87] CLARK, P. und M. MIRMEHDI: *Finding Text Regions Using Localised Measures*. In: *Proceedings of the 11th British Machine Vision Conference*, S. 675–684. BMVA Press, September 2000.
- [88] CLARK, P. und M. MIRMEHDI: *Estimating the orientation and recovery of text planes in a single image*. In: *Proceedings of the 12th British Machine Vision Conference*, S. 421–430. BMVA Press, September 2001.
- [89] COOPER, M., J. FOOTE, A. GIRGENSOHN und L. WILCOX: *Temporal event clustering for digital photo collections*. In: *Proceedings of the 11th ACM international conference on Multimedia*, S. 364–373. ACM Press, 2003.
- [90] COOPER, M. D. und J. FOOTE: *Summarizing video using non-negative similarity matrix factorization*. In: *IEEE Workshop on Multimedia Signal Processing*, S. 25–28. IEEE Computer Society Press, 2002.
- [91] COORG, S., N. MASTER und S. TELLER: *Acquisition of a large pose-mosaic dataset*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 872–878. IEEE Computer Society Press, Juni 1998.
- [92] CORMEN, T. H., C. E. LEISERSON, R. L. RIVEST und C. STEIN: *Introduction to Algorithms*. MIT Press, Cambridge, MA, 2. Aufl., 2001.
- [93] COSTA, L. und R. M. CESAR, JR.: *Shape Analysis and Classification*. CRC Press, Boca Raton, FL, USA, September 2000.
- [94] COTTRELL, G. und M. FLEMING: *Face recognition using unsupervised feature extraction*. In: *International Conference on Neural Network*, S. 322–325, 1990.

- [95] COTTRELL, G. und J. METCALFE: *Face, gender, and emotion recognition using holons*. In: *Advances in neural information processing systems*, Bd. 3, S. 564–571, 1991.
- [96] COTTRELL, G., P. MUNRO und D. ZIPSER: *Learning internal representations from grey-scale images: an example of extensional programming*. In: *Proceedings of 9th Annual Cognitive Science Society Conference*, S. 461–473, 1987.
- [97] COTTRELL, G. W. und P. MUNRO: *Principal component analysis of images via back propagation*. In: *Proceedings of IS&T/SPIE conference on Morphological algorithms for analysis of geological phase structure*, Bd. 1001, S. 1070–1076, Januar 1988.
- [98] COURTNEY, J. D.: *Automatic, object-based indexing for assisted analysis of video data*. In: *Proceedings of ACM international conference on Multimedia*, S. 423–424. ACM Press, 1997.
- [99] CRANDALL, D. und R. KASTURI: *Robust Detection of Stylized Text Events in Digital Video*. In: *Proceedings of International Conference on Document Analysis and Recognition*, S. 865–869, September 2001.
- [100] CRAW, I., H. ELLIS und J. LISHMAN: *Automatic extraction of face features*. In: *Pattern Recognition Letters*, Bd. 5, S. 183–187, 1987.
- [101] CRAW, I., D. TOCK und A. BENNETT: *Finding Face Features*. In: *European Conference on Computer Vision*, S. 92–96, 1992.
- [102] CUI, Y. und Q. HUANG: *Extracting characters of license plates from video sequences*. In: *Machine Vision and Applications*, Bd. 10, S. 308–320, April 1998.
- [103] CUN, Y. L.: *Learning process in an asymmetric threshold network*. In: BIENENSTOCK, E. (Hrsg.): *Disordered Systems and Biological Organization*, Bd. 20 d. Reihe *Computer and Systems Sciences*. Springer Verlag, New York, NY, USA, 1986.
- [104] CURRAN, K. und S. ANNESLEY: *Transcoding media for bandwidth constrained mobile devices*. In: *International Journal of Network Management*, Bd. 15(2), S. 75–88. John Wiley & Sons, Inc., März 2005.
- [105] CUTZU, F. und M. J. TARR: *The representation of three-dimensional object similarity in human vision*. In: *Proceedings of IS&T/SPIE conference on Human Vision and Electronic Imaging II*, Bd. 3016, S. 460–471, 1997.
- [106] DANI, P. und S. CHAUDHURI: *Automated assembling of images: Image montage preparation*. In: *Pattern Recognition*, Bd. 28(3), S. 431–445, März 1995.
- [107] DANIEL, G. und M. CHEN: *Video Visualization*. In: *Proceedings of IEEE Visualization*, S. 409–416. IEEE Computer Society Press, Oktober 2003.
- [108] DANIEL, S., S. GUILLAUMEUX und E. MAILLARD: *Adaptation of a partial shape recognition approach*. In: *IEEE International Conference on Systems, Man, and Cybernetics*, Bd. 3, S. 2157–2162. IEEE Computer Society Press, Oktober 1997.
- [109] DAVIS, J. W. und S. VAKS: *A perceptual user interface for recognizing head gesture acknowledgements*. In: *Proceedings of the 2001 workshop on Perceptive user interfaces*, Bd. 15, S. 1–7. ACM Press, 2001.

- [110] DE MERS, D. und G. COTTRELL: *Non-linear Dimensionality Reduction*. In: *Advances in Neural Information Processing Systems*, Bd. 5, S. 580–587. Morgan Kaufmann, 1993.
- [111] DELAMARRE, Q. und O. FAUGERAS: *3D Articulated Models and Multi-View Tracking with Silhouettes*. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Bd. 2, S. 716–721. IEEE Computer Society, 1999.
- [112] DEMENTHON, D., V. KOBLA und D. DOERMANN: *Video summarization by curve simplification*. In: *Proceedings of the sixth ACM international conference on Multimedia*, S. 211–218. ACM Press, 1998.
- [113] DIAZ, M. E., E. DECENCIÈRE und J. SERRA: *A model-based method for line scratches detection and removal in degraded motion picture sequences*. Techn. Ber. 187, Centre de Morphologie Mathématique, Fontainebleau, 1999.
- [114] DIMITROVA, N., H.-J. ZHANG, B. SHAHRARAY, I. SEZAN, T. HUANG und A. ZAKHOR: *Applications of Video-Content Analysis and Retrieval*. In: *IEEE MultiMedia*, Bd. 9(3), S. 42–55. IEEE Computer Society Press, Juli 2002.
- [115] DIVAKARAN, A., K. A. PEKER, R. RADHARKISHNAN, Z. XIONG und R. CABASSON: *Video Summarization Using MPEG-7 Motion Activity and Audio Descriptors*. In: ROSENFELD, A., D. DOERMANN und D. DEMENTHON (Hrsg.): *Video Mining*, Bd. 6. Kluwer Academic Publishers, Oktober 2003.
- [116] DOERMANN, D., J. LIANG und H. LI: *Progress in Camera-Based Document Image Analysis*. In: *International Conference on Document Analysis and Recognition (ICDAR)*, Bd. 1, S. 606–617, 2003.
- [117] DOMSCHKE, W. und A. DREXL: *Einführung in Operations Research*. Springer Verlag, Berlin, Heidelberg, New York, 6. Aufl., 2004.
- [118] DONATO, G., M. S. BARTLETT, J. C. HAGER, P. EKMAN und T. J. SEJNOWSKI: *Classifying Facial Actions*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 21(10), S. 974–989. IEEE Computer Society Press, Oktober 1999.
- [119] DOR, D. und U. ZWICK: *Selecting the median*. In: *Proceedings of ACM-SIAM symposium on Discrete algorithms*, S. 28–37. Society for Industrial and Applied Mathematics, 1995.
- [120] DOUGHERTY, E. R.: *An Introduction to Morphological Image Processing*. SPIE press, Bellingham, Wash, 1992.
- [121] DREW, M. S. und J. AU: *Video keyframe production by efficient clustering of compressed chromaticity signatures*. In: *Proceedings of the 8th ACM international conference on Multimedia*, S. 365–367. ACM Press, 2000.
- [122] DREW, M. S., Z.-N. LI und X. ZHONG: *Video dissolve and wipe detection via spatio-temporal images of chromatic histogram differences*. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Bd. 3, S. 929–932. IEEE Computer Society Press, 2000.

- [123] EGGLESTON, P.: *Constraint-based feature indexing and retrieval for image databases*. In: *Proceedings of IS&T/SPIE conference on Digital Image Processing and Visual Communications Technologies in the Earth and Atmospheric Sciences II*, Bd. 1819, S. 27–39, 1992.
- [124] EIDENBERGER, H.: *Statistical analysis of content-based MPEG-7 descriptors for image retrieval*. In: *ACM Multimedia Systems*, Bd. 10(2), S. 84–97. Springer, August 2004.
- [125] EKIN, A., A. M. TEKALP und R. MEHROTRA: *Automatic soccer video analysis and summarization*. In: *IEEE Transactions on Image Processing*, Bd. 12(7), S. 796–807. IEEE Computer Society Press, Juli 2003.
- [126] ELLIMAN, D. G. und I. T. LANCASTER: *A review of segmentation and contextual analysis techniques for text recognition*. In: *Pattern Recognition*, Bd. 23 (3-4), S. 337–346, März 1990.
- [127] ENKELMANN, W.: *Investigations of multigrid algorithms for the estimation of optical flow fields in image sequences*. In: *Computer Vision, Graphics, and Image Processing*, Bd. 43, S. 150–177, 1988.
- [128] ER, M. J., S. WU, J. LU und H. L. TOH: *Face recognition with radial basis function (RBF) neural networks*. In: *IEEE Transactions on Neural Networks*, Bd. 13(3), S. 697–710. IEEE Computer Society Press, Mai 2002.
- [129] ESSA, I. A. und A. P. PENTLAND: *Facial expression recognition using a dynamic model and motion energy*. In: *Proceedings of IEEE International Conference on Computer Vision*, S. 360–367. IEEE Computer Society Press, 1995.
- [130] FABLET, R. und P. BOUTHEMY: *Spatio-Temporal Segmentation and General Motion Characterization for Video Indexing and Retrieval*. In: *DELOS Workshop on Audio-Visual Digital Libraries*, S. 1–5, Juni 1999.
- [131] FAIRCHILD, M. D.: *Color Appearance Models*. Wiley-IS&T, Chichester, UK, 2. Aufl., 2005.
- [132] FAN, L. und K. K. SUNG: *Model-based varying pose face detection and facial feature registration in video images*. In: *Proceedings of the 8th ACM international conference on Multimedia*, S. 295–302. ACM Press, 2000.
- [133] FAN, X., X. XIE, W. MA, H. ZHANG und H. ZHOU: *Visual Attention Based Image Browsing on Mobile Devices*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Bd. 1, S. 53–56. IEEE Computer Society Press, Juli 2003.
- [134] FARIN, D.: *Automatic Video Segmentation Employing Object/Camera Modeling*. Doktorarbeit, Technische Universiteit Eindhoven, Eindhoven, The Netherlands, 2005.
- [135] FARIN, D., W. EFFELSBERG und P. H. N. DE WITH: *Robust Clustering-Based Video-Summarization with Integration of Domain-Knowledge*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Bd. 1, S. 89–92. IEEE Computer Society Press, 2002.
- [136] FARIN, D., T. HAENSELMANN, S. KOPF, G. KÜHNE und W. EFFELSBERG: *Segmentation and Classification of Moving Video Objects*. In: FURHT, B. und O. MARQUES (Hrsg.): *Handbook of Video Databases: Design and Applications*, Bd. 8 d. Reihe *Internet and Communications Series*, S. 561–591. CRC Press, Boca Raton, FL, USA, September 2003.

- [137] FAUGERAS, O. D.: *Three-Dimensional Computer Vision : A Geometric Viewpoint*. MIT Press, Cambridge, MA, 2. Aufl., 1996.
- [138] FAYZULLIN, M., V. S. SUBRAHMANIAN, M. ALBANESE und A. PICARIELLO: *The priority curve algorithm for video summarization*. In: *Proceedings of the 2nd ACM international workshop on Multimedia databases*, S. 28–35. ACM Press, 2004.
- [139] FAYZULLIN, M., V. S. SUBRAHMANIAN, A. PICARIELLO und M. L. SAPINO: *The CPR model for summarizing video*. In: *Proceedings of the 1st ACM international workshop on Multimedia databases*, S. 2–9. ACM Press, 2003.
- [140] FELDMAN, A. J. und D. H. BALLARD: *Connectionist models and their properties*. In: *Cognitive Science*, Bd. 6, S. 205–254, 1982.
- [141] FELDMAN, J. A.: *A Connectionist Model of Visual Memory*. In: HINTON, G. E. und J. A. ANDERSON (Hrsg.): *Parallel Models of Associative Memory*, S. 65–97. Erlbaum, Hillsdale, NY, USA, 2. Aufl., 1989.
- [142] FISCHER, S., R. LIENHART und W. EFFELSBERG: *Automatic Recognition of Film Genres*. In: *ACM Multimedia*, S. 295–304. ACM Press, November 1995.
- [143] FLEMING, M. K. und G. W. COTTRELL: *Categorization of faces using unsupervised feature extraction*. In: *Proceeding of International Joint Conference on Neural Networks II*, S. 65–70, 1990.
- [144] FLORIANI, L. D.: *A graph based approach to object feature recognition*. In: *Proceedings of the 3rd annual symposium on Computational geometry*, S. 100–109. ACM Press, 1987.
- [145] FLOYD, R. und L. STEINBERG: *An adaptive algorithm for spatial grey scale*. In: *Journal of the Society for Information Display*, Bd. 17(2), S. 75–77, 1976.
- [146] FORESTI, G. L., C. MICHELONI, L. SNIDARO, P. REMAGNINO und T. ELLIS: *Active video-based surveillance system: the low-level image and video processing techniques needed for implementation*. In: *IEEE Signal Processing Magazine*, Bd. 22(2), S. 25–37. IEEE Computer Society Press, März 2005.
- [147] FOX, A., S. GRIBBLE, Y. CHAWATHE und E. BREWER: *Adapting to Network and Client Variation Using Infrastructural Proxies: Lessons and Perspectives*. In: *IEEE Personal Communication*, Bd. 5(4), S. 10–19. IEEE Computer Society Press, 1998.
- [148] FREEMAN, W. T., P. A. BEARDSLEY, H. KAGE, K.-I. TANAKA, K. KYUMA und C. D. WEISSMAN: *Computer vision for computer interaction*. In: *ACM SIGGRAPH Computer Graphics*, Bd. 33(4), S. 65–68. ACM Press, 1999.
- [149] FROMHERZ, T.: *Face Recognition: a Summary of 1995 – 1997*. Techn. Ber. TR-98-027, Berkeley, Berkeley, CA, USA, 1998.
- [150] FROMHERZ, T., P. STUCKI und M. BICHSEL: *A Survey of Face Recognition*. Techn. Ber. 97.01, University of Zurich, Zurich, Switzerland, 1997.

- [151] FRÖBA, B., A. ERNST und C. KÜBLBECK: *Real-Time Face Detection*. In: *IASTED International Conference on Signal and Image Processing (SIP)*, S. 479–502, 2002.
- [152] FUKUNAGA, K.: *Introduction to statistical pattern recognition*. Academic Press Professional, Inc., San Diego, CA, USA, 2. Aufl., 1990.
- [153] FUSIELLO, A., M. APRILE, R. MARZOTTO und V. MURINO: *Mosaic of a video shot with multiple moving objects*. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Bd. 2, S. 307–310. IEEE Computer Society Press, 2003.
- [154] GAGE, M. und R. S. HAMILTON: *The heat equation shrinking convex plane curves*. In: *Journal of Differential Geometry*, Bd. 23, S. 69–96, 1986.
- [155] GAO, J., R. T. COLLINS, A. G. HAUPTMANN und H. D. WACTLAR: *Articulated Motion Modeling for Activity Analysis*. In: *Conference on Computer Vision and Pattern Recognition Workshop*, S. 20–27, Juni 2004.
- [156] GAO, J., A. G. HAUPTMANN und H. D. WACTLAR: *Combining motion segmentation with tracking for activity analysis*. In: *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, S. 699–704. IEEE Computer Society Press, Mai 2004.
- [157] GAO, J. und J. YANG: *An Adaptive Algorithm for Text Detection from Natural Scenes*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Bd. 2, S. 84–89. IEEE Computer Society Press, Dezember 2001.
- [158] GARCIA, C. und X. APOSTOLIDIS: *Text Detection and Segmentation in Complex Color Images*. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Bd. 4, S. 2326–2330. IEEE Computer Society Press, Juni 2000.
- [159] GARGI, U., D. CRANDALL, S. ANTANI, T. GANDHI, R. KEENER und R. KASTURI: *A system for automatic text detection in video*. In: *International Conference on Document Analysis and Recognition*, S. 29–32, September 1999.
- [160] GAUVAIN, J., L. LAMEL und G. ADDA: *Transcribing Broadcast News for Audio and Video Indexing*. In: *Communications of the ACM*, Bd. Vol. 43(2), S. 64–70. ACM Press, Februar 2000.
- [161] GIRGENSOHN, A., J. BORECZKY, P. CHIU, J. DOHERTY, J. FOOTE, G. GOLOVCHINSKY, S. UCHIHASHI und L. WILCOX: *A semi-automatic approach to home video editing*. In: *Proceedings of the 13th annual ACM symposium on User interface software and technology*, S. 81–89. ACM Press, 2000.
- [162] GIRGENSOHN, A. und J. S. BORECZKY: *Time-Constrained Keyframe Selection Technique*. In: *Multimedia Tools and Applications*, Bd. 11(3), S. 347–358. Kluwer Academic Publishers, 2000.
- [163] GLLAVATA, J., R. EWERTH und B. FREISLEBEN: *Tracking text in MPEG videos*. In: *Proceedings of ACM international conference on Multimedia*, S. 240–243. ACM Press, 2004.
- [164] GOLDMANN, L., M. KARAMAN und T. SIKORA: *Human Body Posture Recognition Using MPEG-7 Descriptors*. In: *Proceedings of IS&T/SPIE conference on Visual Communications and Image Processing (VCIP)*, Bd. 5308, S. 177–188, Januar 2004.

- [165] GOLOMB, B. A., D. T. LAWRENCE und T. J. SEJNOWSKI: *Sexnet: A neural network identifies sex from human faces*. In: *Advances in Neural Information Processing Systems*, Bd. 3, S. 572–577, 1991.
- [166] GONG, Y. und X. LIU: *Generating Optimal Video Summaries*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, S. 1559–1562. IEEE Computer Society Press, 2000.
- [167] GONG, Y. und X. LIU: *Video summarization using singular value decomposition*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Bd. 2, S. 174–180. IEEE Computer Society Press, 2000.
- [168] GONG, Y. und X. LIU: *Summarizing Video By Minimizing Visual Content Redundancies*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, S. 155–158. IEEE Computer Society Press, 2001.
- [169] GONG, Y. und X. LIU: *Video summarization and retrieval using singular value decomposition*. In: *Multimedia Systems*, Bd. 9(2), S. 157–168. Springer-Verlag, 2003.
- [170] GONZALEZ, R. C. und R. E. WOODS: *Digital Image Processing*. Addison-Wesley, Reading, Massachusetts, 1993.
- [171] GORDON, G. G.: *Face Recognition from Frontal and Profile Views*. In: *Proceedings of International Workshop on Automatic Face- and Gesture-Recognition (IWAfGR)*, S. 47–52, 1995.
- [172] GOTTUMUKKAL, R. und V. K. ASARI: *System level design of real time face recognition architecture based on composite PCA*. In: *Proceedings of the 13th ACM Great Lakes symposium on VLSI*, S. 157–160. ACM Press, 2003.
- [173] GOVINDAN, V. K. und A. P. SHIVAPRASAD: *Character recognition - a review*. In: *Pattern Recognition*, Bd. 23 (7), S. 671–683, Juli 1990.
- [174] GOVINDARAJU, V.: *Locating human faces in photographs*. In: *International Journal of Computer Vision*, Bd. 19(2), S. 129–146, 1996.
- [175] GRAF, H., T. CHEN, E. PETAJAN und E. COSATTO: *Locating Faces and Facial Parts*. In: *International Workshop on Automatic Face and Gesture Recognition*, S. 41–46, 1995.
- [176] GRAF, H. P., E. COSATTO, D. GIBBON, M. KOCHSEIN und E. PETAJAN: *Multimodal system for locating heads and faces*. In: *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition*, S. 88–93. IEEE Computer Society Press, 1996.
- [177] GRAYSON, M.: *The heat equation shrinks embedded plane curves to round points*. In: *Journal of Differential Geometry*, Bd. 26, S. 285–314, 1987.
- [178] GROSS, R., S. BAKER, I. MATTHEWS und T. KANADE: *Face Recognition Across Pose and Illumination*. In: LI, S. Z. und A. K. JAIN (Hrsg.): *Handbook of Face Recognition*. Springer Verlag, New York, NY, USA, Juni 2004.
- [179] GROSS, R., I. MATTHEWS und S. BAKER: *Appearance-Based Face Recognition and Light-Fields*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 26(4), S. 449–465. IEEE Computer Society Press, April 2004.

- [180] GUNES, H., M. PICCARDI und T. JAN: *Face and Body Gesture Recognition for a Vision-Based Multimodal Analyzer*. In: *Proceedings of Workshop on Visual Information Processing (VIP)*, Bd. 36, S. 19–28, Juni 2004.
- [181] GUO, G., S. Z. LI und K. CHAN: *Face Recognition by Support Vector Machines*. In: *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, S. 196–201. IEEE Computer Society Press, 2000.
- [182] GUTTA, S., J. HUANG, I. F. IMAM und H. WECHSLER: *Face and Hand Gesture Recognition Using Hybrid Classifiers*. In: *Proceedings of International Conference on Automatic Face and Gesture Recognition (ICAFGR)*, S. 164–169, 1996.
- [183] HABERÄCKER, P.: *Praxis der Digitalen Bildverarbeitung und Mustererkennung*. Carl Hanser, München, Wien, 1995.
- [184] HAMMOUD, R. und R. MOHR: *Interactive tools for constructing and browsing structures for movie films*. In: *Proceedings of the 8th ACM international conference on Multimedia*, S. 497–498. ACM Press, 2000.
- [185] HAMPAPUR, A., T. E. WEYMOUTH und R. JAIN: *Digital Video Segmentation*. In: *Proceedings of ACM Multimedia 1994*, S. 357–364. ACM Press, 1994.
- [186] HAN, R., P. BHAGWAT, R. LAMAIRE, T. MUMMERT, V. PERRET und J. RUBAS: *Dynamic Adaptation in an Image Transcoding Proxy for Mobile WWW Browsing*. In: *IEEE Personal Communication*, Bd. 5(6), S. 8–17. IEEE Computer Society Press, 1998.
- [187] HAN, S. H., K. J. YOON und I.-S. KWEON: *A new technique for shot detection and key frames selection in histogram space*. In: *Workshop on Image Processing and Image Understanding (IPIU)*, S. 1–6, Januar 2000.
- [188] HANJALIC, A. und H. ZHANG: *An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis*. In: *IEEE Transactions on Circuits and Systems for Video Technology*, Bd. 9(8), S. 1280–1289. IEEE Computer Society Press, 1999.
- [189] HARALICK, R. M., S. R. STERNBERG und X. ZHUANG: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. In: *Image analysis using mathematical morphology*, Bd. 9 (4), S. 532–550. IEEE Computer Society Press, Juli 1987.
- [190] HARDENBERG, C. VON und F. BÉRARD: *Bare-hand human-computer interaction*. In: *Proceedings of the 2001 workshop on Perceptive user interfaces*, Bd. 15, S. 1–8. ACM Press, 2001.
- [191] HARITAOGU, I., D. HARWOOD und L. DAVIS: *W4: Who, When, Where, What: A Real Time System for Detecting and Tracking People*. In: *Face and Gesture Recognition Conference*, S. 222–227, 1998.
- [192] HARMON, L. und W. HUNT: *Automatic Recognition of Human Face Profiles*. In: *Computer Graphics and Image Processing*, Bd. 6(2), S. 135–156, 1977.
- [193] HARMON, L., M. KHAN, R. LASH und P. RAMIG: *Machine identification of human faces*. In: *Pattern Recognition*, Bd. 13(2), S. 97–110, 1981.

- [194] HARRIS, C. und M. STEPHENS: *A combined corner and edge detector*. In: *Proceedings of Alvey Vision Conference*, S. 147–151, 1988.
- [195] HARTLEY, R. I. und A. ZISSERMAN: *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2. Aufl., 2004.
- [196] HAUPTMANN, A. G. und M. A. SMITH: *Text, Speech and Vision for Video Segmentation: The Informedia Project*. In: *Proceedings of AAAI Fall Symposium on Computational Models for Integrating Language and Vision*, November 1995.
- [197] HAUPTMANN, A. G. und M. J. WITBROCK: *Story Segmentation and Detection of Commercials in Broadcast News Video*. In: *Advances in Digital Libraries Conference*, S. 168–179, April 1998.
- [198] HE, L. und A. GUPTA: *Exploring benefits of non-linear time compression*. In: *Proceedings of the 9th ACM international conference on Multimedia*, S. 382–391. ACM Press, 2001.
- [199] HE, L., E. SANOCKI, A. GUPTA und J. GRUDIN: *Auto-summarization of audio-video presentations*. In: *Proceedings of ACM international conference on Multimedia*, S. 489–498. ACM Press, 1999.
- [200] HE, L., E. SANOCKI, A. GUPTA und J. GRUDIN: *Comparing presentation summaries: slides vs. reading vs. listening*. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, S. 177–184. ACM Press, 2000.
- [201] HEISELE, B., P. HO und T. POGGIO: *Face Recognition with Support Vector Machines: Global versus Component-based Approach*. In: *Proceedings of International Conference on Computer Vision (ICCV)*, S. 688–694, 2001.
- [202] HITCHCOCK, F. L.: *The Distribution of a Product from Several Sources to Numerous Localities*. In: *Journal of Mathematics and Physics*, Bd. 20, S. 224–230, 1941.
- [203] HJELMAS, E. und B. K. LOW: *Face detection: A survey*. In: *Computer Vision and Image Understanding*, Bd. 83, S. 236–274, 2001.
- [204] HJELSVOLD, R., S. VDAYGIRI und Y. LEAUTE: *Web-based personalization and management of interactive video*. In: *Proceedings of the 10th international conference on World Wide Web*, S. 129–139, 2001.
- [205] HORN, B. K. und B. G. SCHUNCK: *Determining Optical Flow*. Techn. Ber. A.I. Memo No. 572, MIT, 1980.
- [206] HORN, B. K. und B. G. SCHUNCK: *Determining optical flow*. In: *Artificial Intelligence*, Bd. 17, S. 185–203, 1981.
- [207] HORN, B. K. P.: *Robot Vision*. MIT Electrical Engineering and Computer Science, Cambridge, MA, 1986.
- [208] HOSSAIN, M., A. RAHMAN und A. SADDIK: *A Framework for Repurposing Multimedia Content*. In: *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, S. 971–974. IEEE Computer Society Press, Mai 2004.

- [209] HSIEH, W. W. und A. L. CHEN: *Constructing a bowling information system with video content analysis*. In: *Proceedings of ACM international workshop on Multimedia databases*, S. 18–24. ACM Press, 2003.
- [210] HU, J. und A. BAGGA: *Categorizing Images in Web Documents*. In: *IEEE Multimedia*, Bd. 11(1), S. 22–30. IEEE Computer Society Press, Januar 2004.
- [211] HU, J., J. ZHONG und A. BAGGA: *Combined-media video tracking for summarization*. In: *Proceedings of ACM international conference on Multimedia*, S. 502–505. ACM Press, 2001.
- [212] HU, M. K.: *Visual pattern recognition by moment invariants*. In: *IRE Transactions on Information Theory*, Bd. 8, S. 179–187, 1962.
- [213] HUA, K. A. und J. OH: *Detecting video shot boundaries up to 16 times faster*. In: *Proceedings of ACM international conference on Multimedia*, S. 385–387. ACM Press, 2000.
- [214] HUA, X.-S., X.-R. CHEN, L. WENYIN und H.-J. ZHANG: *Automatic Location of Text in Video Frames*. In: *International Workshop on Multimedia Information Retrieval (MIR)*, 2001.
- [215] HUA, X.-S., L. LU und H.-J. ZHANG: *AVE - Automated Home Video Editing*. In: *ACM Multimedia*, S. 490–497. ACM Press, November 2003.
- [216] HUA, X.-S., L. LU und H.-J. ZHANG: *Photo2Video*. In: *Proceedings of the eleventh ACM international conference on Multimedia*, S. 592–593. ACM Press, November 2003.
- [217] HUA, X.-S., L. WENYIN und H.-J. ZHANG: *An Automatic Performance Evaluation Protocol for Video Text Detection Algorithms*. In: *IEEE Transactions on Circuits and Systems for Video Technology*, Bd. 14 (4), S. 498–507. IEEE Computer Society Press, April 2004.
- [218] HUA, X.-S., P. YIN und H.-J. ZHANG: *Efficient Video Text Recognition Using Multiple Frame Integration*. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*. IEEE Computer Society Press, 2002.
- [219] HUANG, J., V. BLANZ und B. HEISELE: *Face Recognition Using Component-Based SVM Classification and Morphable Models*. In: *Proceedings of the 1st International Workshop on Pattern Recognition with Support Vector Machines*, Bd. 2388, S. 334–341. Springer-Verlag, 2002.
- [220] HUANG, Q., Z. LIU, A. ROSENBERG, D. GIBBON und B. SHAHRARAY: *Automated generation of news content hierarchy by integrating audio, video, and text information*. In: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Bd. 6, S. 3025–3028. IEEE Computer Society Press, 1999.
- [221] HÜRST, W., G. GÖTZ und P. JARVERS: *Advanced user interfaces for dynamic video browsing*. In: *Proceedings of the 12th annual ACM international conference on Multimedia*, S. 742–743. ACM Press, 2004.
- [222] IMAI, A., N. SHIMADA und Y. SHIRAI: *3-D Hand Posture Recognition by Training Contour Variation*. In: *Proceedings of International Conference on Automatic Face and Gesture Recognition*, S. 895–900, 2004.

- [223] IRANI, M. und P. ANANDAN: *About Direct Methods*. In: TRIGGS, B., A. ZISSERMAN und R. SZELISKI (Hrsg.): *Proceedings of International Workshop on Vision Algorithms: Theory and Practice*, Bd. 1883, S. 267–277. Springer Berlin, Heidelberg, September 1999.
- [224] IRANI, M., P. ANANDAN, J. BERGEN, R. KUMAR und S. HSU: *Mosaic representations of video sequences and their applications*. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Bd. 8(4), S. 605–611. IEEE Computer Society Press, Mai 1996.
- [225] IRANI, M., S. HSU und P. ANANDAN: *Video compression using mosaic representations*. In: *Signal Processing: Image Communication*, Bd. 5(3), S. 529–552, 1995.
- [226] ISO/IEC: *Information technology – Coding of audio-visual objects – Part 2: Visual*. Techn. Ber. 14496-2, ISO/IEC, 1999.
- [227] ISO/IEC: *Information technology – Multimedia content description interface (MPEG-7) – Part 3: Visual*. Techn. Ber. TR 15938-3, ISO/IEC, 2002.
- [228] ISO/IEC: *Information technology – Multimedia content description interface (MPEG-7) – Part 8: Extraction and use of MPEG-7 descriptions*. Techn. Ber. TR 15938-8, ISO/IEC, 2002.
- [229] ISO/IEC: *MPEG-21 Multimedia Framework – Part 7: Digital Item Adaptation (Final Committee Draft)*. Techn. Ber. N 5845, ISO/IEC, 2003.
- [230] ISO/IEC: *Information technology – Multimedia framework (MPEG-21) – Part 1: Vision, Technologies and Strategy*. Techn. Ber. TR 21000-1, ISO/IEC, 2004.
- [231] ITTI, L., C. KOCH und E. NIEBUR: *A Model of Saliency-Based Visual Attention for Rapid Scene Analysis*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 20(11), S. 1254–1259. IEEE Computer Society Press, November 1998.
- [232] IWASAWA, S., K. EBIHARA, J. OHYA und S. MORISHIMA: *Real-Time Estimation of Human Body Posture from Monocular Thermal Images*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 15–20. IEEE Computer Society, 1997.
- [233] JACUCCI, G., J. KELA und J. PLOMP: *Configuring gestures as expressive interactions to navigate multimedia recordings from visits on multiple projections*. In: *Proceedings of the 3rd international conference on Mobile and ubiquitous multimedia*, Bd. 83, S. 157–164. ACM Press, 2004.
- [234] JEANNIN, S. und M. BOBER: *Description of core experiments for MPEG-7 motion/shape*. Techn. Ber. JTC 1/SC 29/WG 11 MPEG99/N2690, ISO/IEC, 1999.
- [235] JEANNIN, S. und A. DIVAKARAN: *MPEG-7 visual motion descriptors*. In: *IEEE Transactions on Circuits and Systems for Video Technology*, Bd. 11(6), S. 720–724. IEEE Computer Society Press, Juni 2001.
- [236] JEBARA, T., K. RUSSELL und A. PENTLAND: *Mixtures of Eigenfeatures for Real-Time Structure from Texture*. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, S. 128–138. IEEE Computer Society Press, 1998.

- [237] JI, E.-M., H.-S. YOON und Y. J. BAE: *Touring into the picture using hand shape recognition*. In: *Proceedings of the 8th ACM international conference on Multimedia*, S. 388–390. ACM Press, 2000.
- [238] JIANG, H., T. LIN und H. ZHANG: *Video segmentation with the Support of Audio Segmentation and classification*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Bd. 3, S. 1507–1510. IEEE Computer Society Press, Juli 2000.
- [239] JOLLIFFE, I.: *Principal Component Analysis*. Springer Verlag, New York, 1988.
- [240] JONES, M. J. und J. M. REHG: *Statistical color models with application to skin detection*. In: *International Journal of Computer Vision*, Bd. 46(1), S. 81–96. Kluwer Academic Publishers, Januar 2002.
- [241] JONSSON, K., J. KITTLER, Y. P. LI und J. MATAS: *Learning Support Vectors for Face Verification and Recognition*. In: *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, S. 208–213. IEEE Computer Society Press, 2000.
- [242] JOYCE, R. A. und B. LIU: *Temporal segmentation of video using frame and histogram-space*. In: *International Conference on Image Processing*, Bd. 3, S. 941–944, September 2000.
- [243] JOYEUX, L., S. BOUKIR und B. BESSERER: *Film line scratch removal using Kalman filtering and Bayesian restoration*. In: *Proceedings of the 5th IEEE Workshop on Applications of Computer Vision*, S. 8–13. IEEE Computer Society Press, Dezember 2000.
- [244] JOYEUX, L., O. BUISSON, B. BESSERER und S. BOUKIR: *Detection and removal of line scratches in motion picture films*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Bd. 1, S. 548–553. IEEE Computer Society Press, Juni 1999.
- [245] JUELL, P. und R. MARSH: *A Hierarchical Neural Network for Human Face Detection*. In: *Pattern Recognition*, Bd. 29(5), S. 781–787, 1996.
- [246] JUNG, B., T. KWAK, J. SONG und Y. LEE: *Narrative abstraction model for story-oriented video*. In: *Proceedings of the 12th annual ACM international conference on Multimedia*, S. 828–835. ACM Press, 2004.
- [247] JÄHNE, B.: *Digitale Bildverarbeitung*. Springer Verlag, Berlin, Heidelberg, New York, 2. Aufl., 1991.
- [248] JÄHNE, B.: *Digital Image Processing. Concepts, Algorithms, and Scientific Applications*. Springer Verlag, Berlin, Heidelberg, 4. Aufl., 2000.
- [249] KANG, H., T. F. COOTES und C. TAYLOR: *A Comparison of Face Verification Algorithms using Appearance Models*. In: *British Machine Vision Conference (BMVC)*, S. 477–486, September 2002.
- [250] KANG, H.-B.: *Video abstraction techniques for a digital library*. In: SHIH, T. K. (Hrsg.): *Distributed multimedia databases: techniques and applications*, S. 120–132. Idea Group Publishing, 2002.

- [251] KAPOOR, A., Y. QI und R. W. PICARD: *Fully Automatic Upper Facial Action Recognition*. In: *Workshop on IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG)*, S. 195–202. IEEE Computer Society Press, Oktober 2003.
- [252] KARHUNEN, K.: *Über lineare Methoden in der Wahrscheinlichkeitsrechnung*. In: *Annales Academiae Scientiarum Fennicae, Series AI: Mathematica-Physica*, Bd. 37, S. 3–79, 1946.
- [253] KASIK, D.: *Strategies for Consistent Image Partitioning*. In: *IEEE Multimedia*, Bd. 11(1), S. 32–41. IEEE Computer Society Press, Januar 2004.
- [254] KAYA, Y. und K. KOBAYASHI: *A basic study on human face recognition*. In: *Frontiers of Pattern Recognition*, S. 265–289. Academic Press, New York, NY, USA, 1971.
- [255] KIM, C. und J.-N. HWANG: *A fast and robust moving object segmentation in video sequences*. In: *IEEE International Conference on Image Processing*, S. 131–134. IEEE Computer Society Press, Oktober 1999.
- [256] KIM, C. und J.-N. HWANG: *An integrated scheme for object-based video abstraction*. In: *Proceedings of ACM international conference on Multimedia*, S. 303–311. ACM Press, 2000.
- [257] KIM, C. und J.-N. HWANG: *Fast and Automatic Video Object Segmentation and Tracking for Content-Based Applications*. In: *IEEE Transactions on Circuits and Systems for Video Technology (CSVT)*, Bd. 12(2), S. 122–129. IEEE Computer Society Press, Februar 2002.
- [258] KIM, J. D. und H. K. KIM: *Shape descriptor based on multi-layer eigenvector*. Techn. Ber. JTC 1/SC 29/WG 11, ISO/IEC, Lancaster, UK, 1999.
- [259] KIM, J.-G., Y. WANG und S.-F. CHANG: *Content-adaptive Utility-based Video Adaptation*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, S. 281–284. IEEE Computer Society Press, Juli 2003.
- [260] KIM, N. W., E. K. KANG, J. H. IM, T. Y. KIM und J. S. CHOI: *Scene change detection and classification algorithm on compressed video streams*. In: *International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, S. 279–282, Juli 2001.
- [261] KIMIA, B. B. und K. SIDDIQI: *Geometric heat equation and nonlinear diffusion of shapes and images*. In: *Computer Vision and Image Understanding*, Bd. 64(3), S. 305–322, 1996.
- [262] KIMURA, F. und M. SHRIDHAR: *Handwritten numerical recognition based on multiple algorithms*. In: *Pattern Recognition*, Bd. 24 (10), S. 969–983, 1991.
- [263] KING, T., T. BUTTER, M. BRANTNER, S. KOPF, T. HAENSELMANN, A. BISKOP, A. FÄRBER und W. EFFELSBERG: *Distribution of Fingerprints for 802.11-based Positioning Systems*. Techn. Ber. TR-2006-019, Department for Mathematics and Computer Science, University of Mannheim, Dezember 2006.
- [264] KING, T., T. HAENSELMANN, S. KOPF und W. EFFELSBERG: *Overhearing the Wireless Interface for 802.11-based Positioning Systems*. Techn. Ber. TR-2006-018, Department for Mathematics and Computer Science, University of Mannheim, November 2006.

- [265] KING, T., T. HAENSELMANN, S. KOPF und W. EFFELSBERG: *Positionierung mit Wireless-LAN und Bluetooth*. In: *Praxis der Informationsverarbeitung und Kommunikation*, S. 9–17, März 2006.
- [266] KING, T., S. KOPF und W. EFFELSBERG: *A Location System based on Sensor Fusion: Research Areas and Software Architecture*. In: *Proc. of 2. GI/ITG KuVS Fachgespräch 'Ortsbezogene Anwendungen und Dienste'*, S. 28–32, Stuttgart, Germany, Juni 2005.
- [267] KING, T., S. KOPF und W. EFFELSBERG: *Positionserkennung von Studierenden in Hörsälen mit dem Chi-Quadrat-Anpassungstest*. In: *Proc. of 3. GI/ITG KuVS Fachgespräch 'Ortsbezogene Anwendungen und Dienste'*, S. 44–48, Berlin, Germany, September 2006.
- [268] KING, T., S. KOPF, T. HAENSELMANN, C. LUBBERGER und W. EFFELSBERG: *COMPASS: A Probabilistic Indoor Positioning System Based on 802.11 and Digital Compasses*. In: *Proc. of the First ACM International Workshop on Wireless Network Testbeds, Experimental evaluation and Characterization (WiNTECH 2006)*, S. 34–40, Los Angeles, CA, USA, September 2006.
- [269] KIRBY, M. und L. SIROVICH: *Application of the Karhunen-Loève procedure for the characterization of human faces*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 12(1), S. 103–108. IEEE Computer Society Press, 1990.
- [270] KOHONEN, T.: *Associative Memory: A System Theoretical Approach*. Springer Verlag, New York, 1977.
- [271] KOKARAM, A.: *Detection and removal of line scratches in degraded motion picture sequences*. In: *Signal Processing*, Bd. 1, S. 5–8, September 1996.
- [272] KOKARAM, A. C.: *Removal of line artefacts for digital dissemination of archived film and video*. In: *IEEE International Conference on Multimedia Computing and Systems*, Bd. 2, S. 245–249. IEEE Computer Society Press, Juni 1999.
- [273] KOKARAM, A. C., R. DAHYOT, F. PITIE und H. DENMAN: *Simultaneous Luminance and Position Stabilization for Film and Video*. In: *Proceedings of IS&T/SPIE conference on Visual Communications and Image Processing (VCIP)*, Bd. 5022, S. 688–699, Januar 2003.
- [274] KONEN, W. und E. SCHULZE-KRÜGER: *ZN-Face: A System for Access Control Using Automated Face Recognition*. In: *Proceedings of International Workshop on Automatic Face- and Gesture-Recognition (IWAAGR)*, S. 18–23, 1995.
- [275] KOPF, S.: *Verfahren zur Inhaltsadaption von Darstellungselementen*. Techn. Ber. TR-2005-014, Department for Mathematics and Computer Science, University of Mannheim, Germany, 2005.
- [276] KOPF, S., T. HAENSELMANN und W. EFFELSBERG: *Automatic Generation of Video Summaries for Historical Films*. Techn. Ber. TR-04-008, Department for Mathematics and Computer Science, University of Mannheim, Germany, 2004.
- [277] KOPF, S., T. HAENSELMANN und W. EFFELSBERG: *Enhancing Curvature Scale Space Features for Robust Shape Classification*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, S. 478–481. IEEE Computer Society Press, Juli 2005.

- [278] KOPF, S., T. HAENSELMANN und W. EFFELSBERG: *Robust Character Recognition in Low-Resolution Images and Videos*. Techn. Ber. TR-05-002, Department for Mathematics and Computer Science, University of Mannheim, Germany, 2005.
- [279] KOPF, S., T. HAENSELMANN und W. EFFELSBERG: *Shape-based Posture and Gesture Recognition in Videos*. In: *Proceedings of IS&T/SPIE conference on Storage and Retrieval Methods and Applications for Multimedia*, Bd. 5682, S. 114–124, Januar 2005.
- [280] KOPF, S., T. HAENSELMANN, D. FARIN und W. EFFELSBERG: *Automatic Generation of Summaries for the Web*. In: *Proceedings of IS&T/SPIE conference on Storage and Retrieval for Media Databases*, Bd. 5307, S. 417–428, Januar 2004.
- [281] KOPF, S., T. HAENSELMANN, D. FARIN und W. EFFELSBERG: *Automatic Generation of Video Summaries for Historical Films*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Bd. 3, S. 27–30. IEEE Computer Society Press, Juni 2004.
- [282] KOPF, S., T. KING und W. EFFELSBERG: *Improving the Accuracy of GPS*. In: *Proc. of 2. GI/ITG KuVS Fachgespräch 'Ortsbezogene Anwendungen und Dienste'*, Stuttgart, Germany, Juni 2005.
- [283] KOPF, S., T. KING, F. LAMPI und W. EFFELSBERG: *Automatische Kamerasteuerung in Interaktiven Vorlesungen*. In: *Pervasive University im Rahmen der GI Jahrestagung 2006 (PerU2006)*, Dresden, Germany, Oktober 2006.
- [284] KOPF, S., T. KING, F. LAMPI und W. EFFELSBERG: *Video Color Adaptation for Mobile Devices*. In: *Proceedings of the 14th ACM international conference on Multimedia*, S. 963–964. ACM Press, Oktober 2006.
- [285] KOPF, S. und M. KNAUS: *Verbesserung der Qualität von historischen Filmen*. Techn. Ber. TR-2006-001, Department for Mathematics and Computer Science, University of Mannheim, Germany, 2006.
- [286] KOPF, S., F. LAMPI, T. KING und W. EFFELSBERG: *Automatic Scaling and Cropping of Videos for Devices with Limited Screen Resolution*. In: *Proceedings of the 14th ACM international conference on Multimedia*, S. 957–958. ACM Press, Oktober 2006.
- [287] KOPF, S. und A. OERTEL: *Gesichtserkennung in Bildern und Videos mit Hilfe von Eigenfaces*. Techn. Ber. TR-05-008, Department for Mathematics and Computer Science, University of Mannheim, Germany, 2005.
- [288] KOPF, S., N. SCHEELE und W. EFFELSBERG: *The Interactive Lecture: Teaching and Learning Technologies for Large Classrooms*. Techn. Ber. TR-05-001, Department for Mathematics and Computer Science, University of Mannheim, Januar 2005.
- [289] KOPF, S., N. SCHEELE, L. WINSCHERL und W. EFFELSBERG: *Improving Activity and Motivation of Students with Innovative Teaching and Learning Technologies*. In: *Methods and Technologies for Learning*, S. 551–556, Palermo, Italy, April 2005.
- [290] KOTROPOULOS, C. und I. PITAS: *Rule-based face detection in frontal views*. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Bd. 4, S. 2537–2540. IEEE Computer Society Press, April 1997.

- [291] KOTROPOULOS, C., A. TEFAS und I. PITAS: *Frontal face authentication using variants of dynamic link matching based on mathematical morphology*. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*, S. 122–126. IEEE Computer Society Press, Oktober 1998.
- [292] KRAAIJ, W., A. F. SMEATON und P. OVER: *TRECVID 2004 – An Introduction*. In: *TREC Video Retrieval Evaluation Publications (TRECVID)*, S. 1–13, 2004.
- [293] KWON, Y. und N. D. V. LOBO: *Face detection using templates*. In: *Proceedings of International Conference on Pattern Recognition (ICPR)*, S. 764–767, Oktober 1994.
- [294] KÜHNE, G., S. RICHTER und M. BEIER: *Motion-based Segmentation and Contour-based Classification of Video Objects*. In: *Proceedings ACM Multimedia 2001*, S. 41–50. ACM Press, September 2001.
- [295] KÜHNE, G., J. WEICKERT, O. SCHUSTER und S. RICHTER: *A tensor-driven active contour model for moving object segmentation*. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Bd. II, S. 73–76. IEEE Computer Society Press, Oktober 2001.
- [296] LAMPI, F., S. KOPF und W. EFFELSBERG: *Mediale Aufbereitung von Lehrveranstaltungen und ihre automatische Veröffentlichung - Ein Erfahrungsbericht*. In: *Die 4. e-Learning Fachtagung Informatik der Gesellschaft für Informatik (DeLFI 2006)*, Darmstadt, Germany, September 2006.
- [297] LARIMORE, M. G., C. R. JOHNSON und J. R. TREICHLER: *Theory and Design of Adaptive Filters*. Prentice-Hall, New Jersey, 2001.
- [298] LATECKI, L. J. und R. LAKAMPER: *Shape Similarity Measure Based on Correspondence of Visual Parts*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 22(10), S. 1185–1190. IEEE Computer Society Press, 2000.
- [299] LAVIOLA, J. J.: *A survey of hand posture and gesture recognition techniques and technology*. Techn. Ber. CS-99-11, Department of Computer Science, Brown University, Juni 1999.
- [300] LEE, C., S. GHYME, C. PARK und K. WOHN: *The control of avatar motion using hand gesture*. In: *Proceedings of the ACM symposium on Virtual reality software and technology*, S. 59–65. ACM Press, 1998.
- [301] LEI, Z. und N. D. GEORGANAS: *Context-based Media Adaptation in Pervasive Computing*. In: *Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering*, Bd. 2, S. 913–918. IEEE Computer Society Press, Mai 2001.
- [302] LEI, Z. und N. D. GEORGANAS: *Rate adaptation transcoding for precoded video streams*. In: *Proceedings of the 10th ACM international conference on Multimedia*, S. 127–136. ACM Press, 2002.
- [303] LEO, M., T. D’ORAZIO und P. SPAGNOLO: *Human activity recognition for automatic visual surveillance of wide areas*. In: *Proceedings of the ACM 2nd international workshop on Video surveillance and sensor networks*, S. 124–130. ACM Press, 2004.

- [304] LEUNG, T. K., M. C. BURL und P. PERONA: *Finding faces in cluttered scenes using random labeled graph matching*. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, S. 637–644. IEEE Computer Society Press, 1995.
- [305] LEVENBERG, K.: *A Method for the Solution of Certain Non-Linear Problems in Least Squares*. In: *Quarterly of Applied Math.*, Bd. 2, S. 164–168, 1944.
- [306] LI, F. C., A. GUPTA, E. SANOCKI, L. WEI HE und Y. RUI: *Browsing digital video*. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, S. 169–176. ACM Press, 2000.
- [307] LI, H., D. DOERMANN und O. KIA: *Automatic text detection and tracking in digital videos*. In: *IEEE Transactions on Image Processing*, Bd. 9, S. 147–156. IEEE Computer Society Press, Januar 2000.
- [308] LI, Y., S. GONG und H. LIDDELL: *Video-based online face recognition using identity surfaces*. In: *Proceedings of IEEE International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems (RATFG-RTS)*, S. 40–46. IEEE Computer Society Press, Juli 2001.
- [309] LI, Y., W. MING und C.-C. J. KUO: *Semantic video content abstraction based on multiple cues*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, S. 159–162. IEEE Computer Society Press, 2001.
- [310] LI, Y., T. ZHANG und D. TRETTER: *An overview of video abstraction techniques*. Techn. Ber. HPL–2001–191, HP Laboratory, 2001.
- [311] LIE, H. und J. SAARELA: *Multipurpose Web Publishing Using HTML, XML and CSS*. In: *Communications of the ACM*, Bd. 42(10), S. 95–101. ACM Press, Oktober 1999.
- [312] LIEN, J., T. KANADE, J. COHN und C. LI: *Detection, tracking, and classification of subtle changes in facial expression*. In: *Journal of Robotics and Autonomous Systems*, Bd. 31, S. 131–146, 2000.
- [313] LIENHART, R.: *Verfahren zur Inhaltsanalyse, zur Indizierung und zum Vergleich von digitalen Videosequenzen*. Doktorarbeit, University of Mannheim, Mannheim, Germany, 1998.
- [314] LIENHART, R.: *Abstracting home video automatically*. In: *Proceedings of the 7th ACM international conference on Multimedia*, S. 37–40. ACM Press, 1999.
- [315] LIENHART, R.: *Comparison of Automatic Shot Boundary Detection Algorithms*. In: *Proceedings of IS&T/SPIE conference on Video Processing VII*, Bd. 3656, S. 290–301, Januar 1999.
- [316] LIENHART, R.: *Dynamic video summarization of home video*. In: *Proceedings of IS&T/SPIE conference on Storage and Retrieval for Media Databases 2000*, Bd. 3972, S. 378–389, 1999.
- [317] LIENHART, R.: *Reliable Dissolve Detection*. In: *Proceedings of IS&T/SPIE conference on Storage and Retrieval for Media Databases*, Bd. 4315, S. 219–230, 2001.
- [318] LIENHART, R.: *Reliable Transition Detection In Videos: A Survey and Practitioner's Guide*. In: *International Journal of Image and Graphics (IJIG)*, Bd. 1, S. 469–486, 2001.

- [319] LIENHART, R.: *Video OCR: A Survey and Practitioner's Guide*. In: ROSENFELD, A., D. DOERMANN und D. DEMENTHON (Hrsg.): *Video Mining*, Bd. 6. Kluwer Academic Publishers, Oktober 2003.
- [320] LIENHART, R. und W. EFFELSBERG: *Automatic Text Segmentation and Text Recognition for Video Indexing*. In: *ACM/Springer Multimedia Systems*, Bd. 8, S. 69–81. ACM Press, Januar 2000.
- [321] LIENHART, R., W. EFFELSBERG und R. JAIN: *VisualGREP: A Systematic Method to Compare and Retrieve Video Sequences*. In: *Multimedia Tools and Applications*, Bd. 10(1), S. 47–72, 2000.
- [322] LIENHART, R., S. PFEIFFER und W. EFFELSBERG: *Automatic Movie Abstracting*. Techn. Ber. TR-97-003, Department for Mathematics and Computer Science, University of Mannheim, Germany, 1997.
- [323] LIENHART, R., S. PFEIFFER und W. EFFELSBERG: *Video Abstracting*. In: *Communications of the ACM*, Bd. 40, S. 55–62. ACM Press, 1997.
- [324] LIENHART, R. und A. WERNICKE: *Localizing and Segmenting Text in Images and Videos*. In: *IEEE Transactions on Circuits and Systems for Video Technology*, Bd. 12 (4), S. 256–268. IEEE Computer Society Press, April 2002.
- [325] LIN, T. und H.-J. ZHANG: *Automatic Video Scene Extraction by Shot Grouping*. In: *International Conference on Pattern Recognition (ICPR)*, Bd. 4, S. 4039–4042, 2000.
- [326] LITER, J. C., B. S. TJAN, H. H. BÜLTHOFF und N. KÖHNEN: *Viewpoint Effects in Naming Silhouette and Shaded Images of Familiar Objects*. Techn. Ber. 54, Max-Planck-Institut for Biological Cybernetics, Tübingen, Germany, 1997.
- [327] LIU, H., X. XIE, W.-Y. MA und H.-J. ZHANG: *Automatic browsing of large pictures on mobile devices*. In: *Proceedings of the 11th ACM international conference on Multimedia*, S. 148–155. ACM Press, 2003.
- [328] LONCARIC, S.: *A Survey of Shape Analysis Techniques*. In: *Pattern Recognition*, Bd. 31(8), S. 983–1001, August 1998.
- [329] LOWE, D. G.: *Distinctive Image Features from Scale-Invariant Keypoints*. In: *International Journal of Computer Vision*, Bd. 60(2), S. 91–110. Kluwer Academic Publishers, November 2004.
- [330] LOÈVE, M. M.: *Probability Theory*. Van Nostrand, Princeton, N.J., 1955.
- [331] LU, C., M. S. DREW und J. AU: *Classification of summarized videos using hidden markov models on compressed chromaticity signatures*. In: *Proceedings of the 9th ACM international conference on Multimedia*, S. 479–482. ACM Press, 2001.
- [332] LU, J., K. PLATANIOTIS und A. VENETSANOPOULOS: *A Kernel Machine Based Approach For Multi-view Face Recognition*. In: *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Bd. 1, S. 265–268. IEEE Computer Society Press, September 2002.

- [333] LU, X.: *Image Analysis for Face Recognition – A brief survey*. Techn. Ber. 1, Computer Science and Engineering, Michigan State University, 2003.
- [334] LU, X., D. COLBRY und A. K. JAIN: *Three-Dimensional Model Based Face Recognition*. In: *Proceedings of International Conference on Pattern Recognition (ICPR)*, Bd. 1, S. 362–366, August 2004.
- [335] LUM, W. und F. LAU: *A Context-Aware Decision Engine for Content Adaptation*. In: *IEEE Pervasive Computing*, Bd. 1(3), S. 41–49. IEEE Computer Society Press, Juli 2002.
- [336] LUO, H. und A. ELEFThERiADiS: *On face detection in the compressed domain*. In: *Proceedings of the 8th ACM international conference on Multimedia*, S. 285–294. ACM Press, 2000.
- [337] LUO, H. und J. FAN: *Concept-oriented video skimming and adaptation via semantic classification*. In: *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, S. 213–220. ACM Press, 2004.
- [338] MA, H. und D. S. DOERMANN: *Adaptive Hindi OCR using generalized Hausdorff image comparison*. In: *ACM Transactions on Asian Language Information Processing (TALIP)*, Bd. 2 (3), S. 193–218. ACM Press, 2003.
- [339] MA, Y.-F., L. LU, H.-J. ZHANG und M. LI: *A User Attention Model for Video Summarization*. In: *Proceedings of the 10th ACM international conference on Multimedia*, S. 533–542. ACM Press, 2002.
- [340] MA, Y.-F., J. SHENG, Y. CHEN und H.-J. ZHANG: *MSR-Asia at TREC-10 Video Track: Shot Boundary Detection Task*. In: *Text Retrieval Conference (TREC) – Video Track*, S. 142–150, 2001.
- [341] MACIEL, J. und J. P. COSTEIRA: *A Global Solution to Sparse Correspondence Problems*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 25(2), S. 187–199. IEEE Computer Society, Februar 2003.
- [342] MACKWORTH, A. K. und F. MOKHTARIAN: *Scale-Based Descriptions of Planar Curves*. In: *Proceedings of Canadian Society for Computational Studies of Intelligence*, S. 114–119, 1984.
- [343] MACKWORTH, A. K. und F. MOKHTARIAN: *The renormalized curvature scale space and the evolution properties of planar curves*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 318–326. IEEE Computer Society Press, 1988.
- [344] MALIK, S. und J. LASZLO: *Visual touchpad: a two-handed gestural input device*. In: *Proceedings of the 6th international conference on Multimodal interfaces (ICMI)*, S. 289–296. ACM Press, 2004.
- [345] MANTAS, J.: *An Overview Of Character Recognition Methodologies*. In: *Pattern Recognition*, Bd. 19, S. 425–430, 1986.
- [346] MARQUARDT, D. W.: *An Algorithm for Least-Squares Estimation of Nonlinear Parameters*. In: *J. Soc. Indust. Appl. Math.*, Bd. 11(2), S. 431–441, 1963.

- [347] MARR, D.: *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, San Francisco, CA, USA, 1982.
- [348] MARR, D. und E. HILDRETH: *Theory of edge detection*. In: *Proceedings of the Royal Society of London, Series B*, Bd. 270, S. 187–217, 1980.
- [349] MAURER, T. und C. VON DER MALSBURG: *Single-View Based Recognition of Faces Rotated in Depth*. In: *Proceedings of International Workshop on Automatic Face- and Gesture-Recognition (IWAAGR)*, S. 248–253, 1995.
- [350] MCMILLAN, L. und G. BISHOP: *Plenoptic modeling: An image-based rendering system*. In: *Proceedings of Computer graphics and interactive techniques*, S. 39–46. ACM Press, 1995.
- [351] MENTZELOPOULOS, M. und A. PSARROU: *Key-frame extraction algorithm using entropy difference*. In: *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, S. 39–45. ACM Press, 2004.
- [352] MERIALDO, B., K. T. LEE, D. LUPARELLO und J. ROUDAIRE: *Automatic construction of personalized TV news programs*. In: *Proceedings of the 7th ACM international conference on Multimedia*, S. 323–331. ACM Press, 1999.
- [353] MIAO, J., B. YIN, K. WANG, L. SHEN und X. CHEN: *A hierarchical multiscale and multiangle system for human face detection in a complex background using gravity-center template*. In: *Pattern Recognition*, Bd. 32(7), S. 1237–1248, 1999.
- [354] MIENE, A., A. DAMMEYER, T. HERMES und O. HERZOG: *Advanced and Adaptive Shot Boundary Detection*. In: *Proceedings of European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, September 2001.
- [355] MIKOLAJCZYK, K. und C. SCHMID: *A Performance Evaluation of Local Descriptors*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 27(10), S. 1615–1630. IEEE Computer Society, Oktober 2005.
- [356] MILGRAM, D. L.: *Computer methods for creating photomosaics*. In: *IEEE Transactions on Computers*, Bd. C–24, S. 1113–1119. IEEE Computer Society Press, 1975.
- [357] MOESLUND, T. B. und E. GRANUM: *3D Human Pose Estimation using 2D-Data and an Alternative Phase Space Representation*. In: *IEEE Workshop on Human Modeling, Analysis and Synthesis*, S. 26–33. IEEE Computer Society, Juni 2000.
- [358] MOGHADDAM, B., C. NASTAR und A. PENTLAND: *Bayesian Face Recognition using Deformable Intensity Surfaces*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 638–645. IEEE Computer Society Press, Juni 1996.
- [359] MOHAN, R., J. SMITH und C. LI: *Adapting Multimedia Internet Content For Universal Access*. In: *IEEE Transactions on Multimedia*, Bd. 1(1), S. 104–114. IEEE Computer Society Press, März 1999.
- [360] MOKHTARIAN, F.: *Silhouette-Based Isolated Object Recognition through Curvature Scale Space*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 17(5), S. 539–544. IEEE Computer Society Press, 1995.

- [361] MOKHTARIAN, F.: *A Theory of Multi-Scale, Torsion-Based Shape Representation for Space Curves*. In: *Computer Vision and Image Understanding*, Bd. 68 (1), S. 1–17, 1997.
- [362] MOKHTARIAN, F., S. ABBASI und J. KITTLER: *Efficient and Robust Retrieval by Shape Content through Curvature Scale Space*. In: *Proceedings of International Workshop on Image Databases and Multimedia Search*, S. 35–42, 1996.
- [363] MOKHTARIAN, F., S. ABBASI und J. KITTLER: *Robust and Efficient Shape Indexing through Curvature Scale Space*. In: *British Machine Vision Conference*, 1996.
- [364] MOKHTARIAN, F. und M. BOBER: *Curvature Scale Space Representation: Theory, Applications, and MPEG-7 Standardization (Computational Imaging and Vision, 25)*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.
- [365] MOKHTARIAN, F. und A. K. MACKWORTH: *Scale-Based Description and Recognition of Planar Curves and Two-Dimensional Shapes*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 8(1), S. 34–43. IEEE Computer Society Press, 1986.
- [366] MOKHTARIAN, F. und A. K. MACKWORTH: *A Theory of Multiscale, Curvature-Based Shape Representation for Planar Curves*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 14(8), S. 789–805. IEEE Computer Society Press, August 1992.
- [367] MORI, G., S. BELONGIE und J. MALIK: *Shape contexts enable efficient retrieval of similar shapes*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Bd. 1, S. 723–730. IEEE Computer Society Press, 2001.
- [368] MORIYAMA, T. und M. SAKAUCHI: *Video summarisation based on the psychological content in the track structure*. In: *Proceedings of the 2000 ACM workshops on Multimedia*, S. 191–194. ACM Press, 2000.
- [369] MU, X. und G. MARCHIONINI: *Statistical visual feature indexes in video retrieval*. In: *Proceedings of ACM SIGIR conference on Research and Development in Informaion Retrieval*, S. 395–396. ACM Press, 2003.
- [370] MULHEM, P., J. GENSEL und H. MARTIN: *Adaptive video summarization*. In: FURHT, B. und O. MARQUES (Hrsg.): *Handbook of Video Databases: Design and Applications*, Bd. 8 d. Reihe *Internet and Communications Series*, S. 279–298. CRC Press, Boca Raton, FL, USA, September 2003.
- [371] MURAKAMI, K. und H. TAGUCHI: *Gesture recognition using recurrent neural networks*. In: *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, S. 237–242. ACM Press, 1991.
- [372] MYERS, G., R. BOLLES, Q.-T. LUONG und J. HERSON: *Recognition of Text in 3-D Scenes*. In: *4th Symposium on Document Image Understanding Technology*, S. 23–25, April 2001.
- [373] MÜLLER, D.: *Automatische Detektion von Gesichtern in Bewegtbildern*. Diplomarbeit, University of Mannheim, Mannheim, Germany, März 1997.

- [374] NAGAO, K., S. OHIRA und M. YONEOKA: *Annotation-based multimedia summarization and translation*. In: *Proceedings of the 19th international conference on Computational linguistics*, Bd. 1, S. 1–7, 2002.
- [375] NAGY, G.: *At the frontiers of OCR*. In: *Proceedings of the IEEE*, Bd. 80 (7), S. 1093–1100. IEEE Computer Society Press, Juli 1992.
- [376] NAGY, G., T. A. NARTKER und S. V. RICE: *Optical Character Recognition: An Illustrated Guide to the Frontier*. In: *Proceedings of IS&T/SPIE conference on Document Recognition and Retrieval VII*, Bd. 3967, S. 58–69, 2000.
- [377] NAM, J. und A. H. TEWFIK: *Dynamic video summarization and visualization*. In: *Proceedings of the 7th ACM international conference on Multimedia*, S. 53–56. ACM Press, 1999.
- [378] NANG, J., J. JEONG, S. PARK und H. CHA: *An Abstraction of Low Level Video Features for Automatic Retrievals of Explosion Scenes*. In: *IEEE Pacific Rim Conference on Multimedia 2002*, S. 200–208. Springer Verlag, 2002.
- [379] NEPAL, S. und U. SRINIVASAN: *DAVE: A System for Quality Driven Adaptive Video Delivery*. In: *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, S. 223–230. ACM Press, 2003.
- [380] NEUMANN, K. und M. MORLOCK: *Operations Research*. Carl Hanser, München, Wien, 2. Aufl., 2002.
- [381] NG, T. D., H. D. WACTLAR, A. G. HAUPTMANN und M. G. CHRISTEL: *Collages as Dynamic Summaries of Mined Video Content for Intelligent Multimedia Knowledge Management*. In: *AAAI Spring Symposium Series on Intelligent Multimedia Knowledge Management*, März 2003.
- [382] NGO, C. W.: *A Robust Dissolve Detector by Support Vector Machine*. In: *Proceedings of ACM Multimedia Conference*, S. 283–286. ACM Press, 2003.
- [383] NGO, C. W. und C. K. CHAN: *Video Text Detection and Segmentation for Optical Character Recognition*. In: *Multimedia Systems*, Bd. 10 (3), S. 261–272, März 2005.
- [384] NGO, C. W., T. C. PONG und R. T. CHIN: *Detection of Gradual Transitions through Temporal Slice Analysis*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Bd. 1, S. 1036–1041. IEEE Computer Society Press, 1999.
- [385] NIBLACK, W., R. BARBER, W. EQUITZ, M. FLICKNER, E. GLASMAN, D. PETKOVIC, P. YANKER, C. FALOUTSOS und G. TABUIN: *QBIC Project: Querying Images By Content Using Color, Texture, and Shape*. In: *Proceedings of IS&T/SPIE conference on Storage and Retrieval for Image and Video Databases*, Bd. 1908, S. 173–187, 1993.
- [386] NISHINO, H., K. UTSUMIYA, D. KURAOKA, K. YOSHIOKA und K. KORIDA: *Interactive two-handed gesture interface in 3D virtual environments*. In: *Proceedings of the ACM symposium on Virtual reality software and technology*, S. 1–8. ACM Press, 1997.
- [387] NOBLE, B., M. SATYANARAYANAN, D. NARAYANAN, J. E. TILTON, J. FLINN, und K. R. WALKER: *Agile Application-Aware Adaptation for Mobility*. In: *Proceedings of the 16th Symposium on Operating System Principles*, S. 276–287, 1997.

- [388] NURNETT, I.: *MPEG-21: Goals and Archievements*. In: *IEEE Multimedia*, Bd. 10(6), S. 60–70. IEEE Computer Society Press, Oktober 2003.
- [389] NÖLKER, C. und H. RITTER: *Visual recognition of continuous hand postures*. In: *IEEE Transactions on Neural Networks*, Bd. 13(4), S. 983–994. IEEE Computer Society Press, Juli 2002.
- [390] OBRENOVIC, Z., D. STARCEVIC und B. SELIC: *A Model-Driven Approach to Content Repurposing*. In: *IEEE Multimedia*, Bd. 11(1), S. 62–71. IEEE Computer Society Press, Januar 2004.
- [391] OERTEL, A.: *Gesichtserkennung in Videos mithilfe von Eigenfaces*. Diplomarbeit, University of Mannheim, Mannheim, Germany, August 2004.
- [392] OH, J. und K. A. HUA: *Efficient and cost-effective techniques for browsing and indexing large video databases*. In: *Proceedings of the 2000 ACM SIGMOD international conference on Management*, S. 415–426. ACM Press, 2000.
- [393] OH, J. und K. A. HUA: *An Efficient Technique for Summarizing Videos using Visual Contents*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, S. 1167–1170. IEEE Computer Society Press, Juli 2000.
- [394] OH, J., K. A. HUA und N. LIANG: *A Content-based Scene Change Detection and Classification Technique using Background Tracking*. In: *Proceedings of IS&T/SPIE conference on Multimedia Computing and Networking*, Bd. 3969, S. 254–265, Januar 2000.
- [395] OH, J., M. THENNERU und N. JIANG: *Hierarchical video indexing based on changes of camera and object motions*. In: *Proceedings of ACM symposium on Applied Computing*, S. 917–921. ACM Press, 2003.
- [396] OH, J.-H., Q. WEN, J.-K. LEE und S. HWANG: *Video Abstraction*. In: DEB, S. (Hrsg.): *Video Data Management and Information Retrieval*, S. 321–346. Idea Group Inc., IIRM Press, 2005.
- [397] OHYA, J.: *Face/gesture analysis/synthesis technologies for human-to-human communications through virtual environments*. In: *Proceedings of the sixth ACM international conference on Multimedia: Face/gesture recognition and their applications*, S. 12–19. ACM Press, 1998.
- [398] OLIVER, N., F. BERARD und A. PENTLAND: *LAFER: Lips and face tracker*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 123–129. IEEE Computer Society Press, 1996.
- [399] OMOIGUI, N., L. HE, A. GUPTA, J. GRUDIN und E. SANOCKI: *Time-compression: systems concerns, usage, and benefits*. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, S. 136–143. ACM Press, 1999.
- [400] O'TOOLE, A. und H. ABDI: *Advances in Cognitive Sciences*, Kap. Connectionist approaches to visually-based feature extraction. Wiley, London, 1989.
- [401] O'TOOLE, A. J., H. ABDI, K. A. DEFFENBACHER und D. VALENTIN: *Low-dimensional representation of faces in higher dimensions of the face space*. In: *Journal of American Optical Society*, Bd. 10, S. 405–411, 1993.

- [402] O'TOOLE, A. J., H. H. BÜLTHOFF, N. F. TROJE und T. VETTER: *Face Recognition across Large Viewpoint Changes*. In: *Proceedings of International Workshop on Automatic Face- and Gesture-Recognition (IWAFGR)*, S. 59–64, 1995.
- [403] OVER, P., T. IANEVA, W. KRAAIJ und A. F. SMEATON: *TRECVID 2005 An Overview*. In: *TREC Video Retrieval Evaluation Proceedings*, S. 1–27. National Institute of Standards and Technology (NIST), März 2006.
- [404] PALMER, S., E. ROSCH und P. CHASE: *Canonical perspective and the perception of objects*. In: LONG, J. und A. BADDELEY (Hrsg.): *Attention and Performance IX*, S. 135–151. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, USA, 1981.
- [405] PAN, Z. und C.-W. NGO: *Structuring home video by snippet detection and pattern parsing*. In: *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, S. 69–76. ACM Press, 2004.
- [406] PARAGIOS, N. und R. DERICHE: *Geodesic active contours and level sets for the detection and tracking of moving objects*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 22(3), S. 266–280. IEEE Computer Society Press, März 2000.
- [407] PARAGIOS, N. und R. DERICHE: *Geodesic Active Regions: A New Paradigm to Deal with Frame Partition Problems in Computer Vision*. In: *Journal of Visual Communication and Image Representation, Special Issue on Partial Differential Equations in Image Processing, Computer Vision and Computer Graphics*, Bd. 13(1), S. 249–268, März 2002.
- [408] PARAGIOS, N. und R. DERICHE: *Geodesic Active Regions and Level Set Methods for Motion Estimation and Tracking*. In: *Computer Vision and Image Understanding*, Bd. 97 (3), S. 259–282. Elsevier Inc., März 2005.
- [409] PARAGIOS, N. und G. TZIRITAS: *Adaptive Detection and Localization of Moving Objects in Image Sequences*. In: *Signal Processing: Image Communication*, Bd. 14 (4), S. 277–296, 1999.
- [410] PARK, J., J. SEO, D. AN und S. CHUNG: *Detection of Human Faces using Skin Color and Eyes*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Bd. 1, S. 133–136. IEEE Computer Society Press, Juli 2000.
- [411] PARKER, D. B.: *A Comparison of Algorithms for Neuron-Like Cells*. In: DENKER, J. (Hrsg.): *Neural Networks for Computing*, S. 327–332. American Institute of Physics, New York, NY, USA, 1986.
- [412] PARSHIN, V. und L. CHEN: *Video Summarization Based on User-defined Constraints and Preferences*. In: *Proceedings of RIAO International Conference*, S. 18–24, 2004.
- [413] PAVLIDIS, T.: *A Review of Algorithms for Shape Analysis*. In: *Computer Graphics and Image Processing*, Bd. 7(2), S. 243–258, April 1978.
- [414] PENG, Y. und C.-W. NGO: *Clip-based similarity measure for hierarchical video retrieval*. In: *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, S. 53–60. ACM Press, 2004.

- [415] PENTLAND, A.: *Perceptual user interfaces: perceptual intelligence*. In: *Communications of the ACM*, Bd. 43(3), S. 35–44. ACM Press, März 2000.
- [416] PERLMUTTER, K., N. CHADDHA, J. BUCKHEIT, R. GRAY und R. OLSHEN: *Text segmentation in mixed-mode images using classification trees and transform tree-structured vector quantization*. In: *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Bd. 4, S. 2231–2234. IEEE Computer Society Press, 1996.
- [417] PFEIFFER, S., R. LIENHART und W. EFFELSBERG: *Scene Determination Based on Video and Audio Features*. In: *Multimedia Tools and Applications*, Bd. 15(1), S. 59–81. Kluwer Academic Publishers, September 2001.
- [418] PFEIFFER, S., R. LIENHART, S. FISCHER und W. EFFELSBERG: *Abstracting Digital Movies Automatically*. In: *Journal of Visual Communication and Image Representation*, Bd. 7, S. 345–353, 1996.
- [419] PFEIFFER, S., R. LIENHART, G. KÜHNE und W. EFFELSBERG: *The MoCA Project – Movie Content Analysis Research at the University of Mannheim*. In: *Informatik '98: Informatik zwischen Bild und Sprache*, 28. Jahrestagung der Gesellschaft für Informatik, S. 329–338, September 1998.
- [420] PHILLIPS, P. J. und Y. VARDI: *Data-Driven Methods in Face Recognition*. In: *Proceedings of International Workshop on Automatic Face- and Gesture-Recognition (IWAAGR)*, S. 65–69, 1995.
- [421] PONCELEON, D., S. SRINIVASAN, A. AMIR, D. PETKOVIC und D. DIKLIC: *Key to effective video retrieval: effective cataloging and browsing*. In: *Proceedings of the sixth ACM international conference on Multimedia*, S. 99–107. ACM Press, 1998.
- [422] PORTER, S. V., M. MIRMEHDI und B. T. THOMAS: *Detection and classification of shot transitions*. In: *Proceedings of British Machine Vision Conference*, S. 73–82. BMVA Press, September 2001.
- [423] PRESS, W. H., S. A. TEUKOLSKY, W. T. VETTERLING und B. P. FLANNERY: *Numerical Recipes in C : The Art of Scientific Computing*. Cambridge University Press, New York, 1992.
- [424] PROPP, M. und A. SAMAL: *Artificial Neural Network Architectures for Human Face Detection*. In: *Intelligent Eng. Systems through Artificial Neural Networks*, Bd. 2, S. 535–540, 1992.
- [425] RADHAKRISHNAN, R., A. DIVAKARAN und Z. XIONG: *A time series clustering based framework for multimedia mining and summarization using audio features*. In: *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, S. 157–164. ACM Press, 2004.
- [426] RASHEED, Z., Y. SHEIKH und M. SHAH: *Semantic Film Preview Classification Using Low-Level Computable Features*. In: *Proceedings of International Workshop on Multimedia Data and Document Engineering (MDDE)*, S. 1–8, September 2003.
- [427] RASHEED, Z., Y. SHEIKH und M. SHAH: *On the use of Computable Features for Film Classification*. In: *IEEE Transactions on Circuits and Systems for Video Technology*, Bd. 15(1), S. 52–64. IEEE Computer Society Press, 2005.

- [428] REN, L., G. SHAKHNAROVICH, J. K. HODGINS, H. PFISTER und P. VIOLA: *Learning silhouette features for control of human motion*. In: *ACM Transactions on Graphics (TOG)*, Bd. 24(4), S. 1303–1331. ACM Press, Oktober 2005.
- [429] RICHTER, S., G. KÜHNE und O. SCHUSTER: *Contour-based Classification of Video Objects*. In: *Proceedings of IS&T/SPIE conference on Storage and Retrieval for Media Databases*, Bd. 4315, S. 608–618, Januar 2001.
- [430] RIST, T. und P. BRANDMEIER: *Customizing Graphics for Tiny Displays of Mobile Devices*. In: *Personal and Ubiquitous Computing*, Bd. 6(4), S. 260–268. Springer, 2002.
- [431] RIST, T. und P. BRANDMEIER: *Customizing Graphics for Tiny Displays of Mobile Devices*. In: *Proceedings of 3rd International Workshop on Human Computer Interaction with Mobile Devices*, S. 1–4, September 2001.
- [432] ROUSSEEUW, P. J. und A. M. LEROY: *Robust Regression and Outlier Detection*. John Wiley, New York, 1987.
- [433] ROUSSEEUW, P. J. und K. VAN DRIESEN: *Computing LTS Regression for Large Data Sets*. In: *Institute of Mathematical Statistics Bulletin*, Bd. 27(6), November/Dezember 1998.
- [434] ROWE, N.: *Content Repurposing for Small Devices*. In: PAGANI, M. (Hrsg.): *Encyclopedia of Multimedia Technology and Networking (Volume I)*, Bd. 1, S. 110–115. The Idea Group, Hershey, PA, USA, April 2005.
- [435] ROWLEY, H., S. BALUJA und T. KANADE: *Human Face Detection in Visual Scenes*. Techn. Ber. CMU-CS-95-158R, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA, 1995.
- [436] ROWLEY, H., S. BALUJA und T. KANADE: *Rotation Invariant Neural Network-Based Face Detection*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society Press, 1998.
- [437] ROWLEY, H. A., S. BALUJA und T. KANADE: *Human Face Detection in Visual Scenes*. In: TOURETZKY, D. S., M. C. MOZER und M. E. HASSELMO (Hrsg.): *Advances in Neural Information Processing Systems*, Bd. 8, S. 875–881. The MIT Press, 1996.
- [438] ROWLEY, H. A., S. BALUJA und T. KANADE: *Neural Network-Based Face Detection*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 20(1), S. 23–38. IEEE Computer Society Press, 1998.
- [439] RUBNER, Y.: *Perceptual Metrics for Image Database Navigation*. Techn. Ber. CS-TR-99-1621, Stanford University, 1999.
- [440] RUBNER, Y. und C. TOMASI: *Perceptual Metrics for Image Database Navigation*, Bd. 594 d. Reihe *Kluwer International Series in Engineering and Computer Science*. Kluwer Academic Publishers, Boston, MA, USA, 2001.
- [441] RUMELHART, D. E., G. E. HINTON und R. J. WILLIAMS: *Learning representations by back-propagating errors*. In: *Nature*, Bd. 323, S. 533–536, 1986.

- [442] RURAINSKY, J. und P. EISERT: *Template-based Eye and Mouth Detection for 3D Video Conferencing*. In: *International Workshop on Very Low Bitrate Video (VLBV)*, S. 23–31, September 2003.
- [443] SABER, E. und A. M. TEKALP: *Frontal-view face detection and facial feature extraction using color, shape, and symmetry based cost functions*. In: *Pattern Recognition Letters*, Bd. 19(8), S. 669–680. Elsevier Science Inc., Juni 1998.
- [444] SAITO, T., T. KOMATSU, T. HOSHI und T. OHUCHI: *Image Processing for Restoration of Old Film Sequences*. In: *Proceedings of 10th International Conference on Image Analysis and Processing*, S. 709–714, 1999.
- [445] SAKAI, T., M. NAGAI und S. FUJIBAYASHI: *Line Extraction and Pattern Detection in a Photograph*. In: *Pattern Recognition*, Bd. 1, S. 233–248, 1969.
- [446] SAKAI, T., M. NAGAO und M. KIDODE: *Processing of Multilevel Pictures by Computer – The Case of Photographs of Human Face*. In: *Systems Computers Controls*, Bd. 2(3), S. 47–54, 1971.
- [447] SAMAL, A. und P. A. IYENGAR: *Automatic recognition and analysis of human faces and facial expressions: a survey*. In: *Pattern Recognition*, Bd. 25(1), S. 65 – 77. Elsevier Science Inc., Januar 1992.
- [448] SAND, P., L. McMILLAN und J. POPOVIC: *Continuous capture of skin deformation*. In: *ACM Transactions on Graphics (TOG)*, Bd. 22(3), S. 578–586. ACM Press, Juli 2003.
- [449] SATO, T., T. KANADE, E. K. HUGHES und M. A. SMITH: *Video OCR for Digital News Archives*. In: *IEEE International Workshop on Content-Based Access of Image and Video Databases (CAIVD)*, S. 52–60. IEEE Computer Society Press, 1998.
- [450] SATO, T., T. KANADE, E. K. HUGHES, M. A. SMITH und S. SATOH: *Video OCR: Indexing digital news libraries by recognition of superimposed captions*. In: *ACM/Springer Multimedia Systems*, Bd. 7, S. 385–395. ACM Press, 1999.
- [451] SAVINO, P.: *Building an Audio-visual Digital Library of Historical Documentaries: The ECHO Project*. In: *D-Lib Magazine*, Bd. 6 (11), S. 3–4, November 2000.
- [452] SAVINO, P. und C. THANOS: *ECHO – European CHronicles On-line*. In: *Cultivate Interactive*, Bd. 1, S. 1–6, Juli 2000.
- [453] SAWHNEY, H. und R. KUMAR: *True multi-image alignment and its application to mosaicing and lens distortion correction*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Bd. 21(3), S. 450–456. IEEE Computer Society Press, 1997.
- [454] SCASSELLATI, B., S. ALEXOPOULOS und M. FLICKNER: *Retrieving images by 2D shape: a comparison of computation methods with human perceptual judgments*. In: *Proceedings of IS&T/SPIE conference on Storage and Retrieval for Image and Video Databases II*, Bd. 2185, S. 2–14, 1994.

- [455] SCHALLAUER, P., A. PINZ und W. HAAS: *Automatic Restoration for 35mm Film*. In: *Journal of Computer Vision Research*, Bd. 1(3), S. 60–85. MIT Press, 1999.
- [456] SCHMID, C., R. MOHR und C. BAUCKHAGE: *Evaluation of Interest Point Detectors*. In: *International Journal of Computer Vision: Special issue on visual surveillance*, Bd. 37(2), S. 151–172. Kluwer Academic Publishers, Juni 2000.
- [457] SERRA, J.: *Image Analysis and Mathematical Morphology – Part II*. Academic Press, New York, 1988.
- [458] SHANABLEH, T. und M. GHANBARI: *Heterogeneous video transcoding to lower spatio-temporal resolution and different encoding formats*. In: *IEEE Transactions on Multimedia*, Bd. 2(2), S. 101–110. IEEE Computer Society Press, Juni 2000.
- [459] SHIPMAN, F., A. GIRGENSOHN und L. WILCOX: *Generation of interactive multi-level video summaries*. In: *Proceedings of the 11th ACM international conference on Multimedia*, S. 392–401. ACM Press, 2003.
- [460] SIMONCELLI, E. P., E. H. ADELSON und D. J. HEEGER: *Probability Distributions of Optical Flow*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 310–315. IEEE Computer Society Press, 1991.
- [461] SIROVICH, L. und M. KIRBY: *Low-dimensional procedure for the characterization of human faces*. In: *Journal of the Optical Society of America A*, Bd. 4(3), S. 519–524, 1987.
- [462] SMEATON, A., J. GILVARRY, G. GORMLEY, B. TOBIN, S. MARLOW und M. MURPHY: *An Evaluation of Alternative Techniques for Automatic Detection of Shot Boundaries in Digital Video*. In: *Proceedings of Irish Machine Vision and Image Processing Conference (IMVIP)*, September 1999.
- [463] SMITH, M. und T. KANADE: *Video Skimming for Quick Browsing Based on Audio and Image Characterization*. Techn. Ber. CMU-CS-95-186, Carnegie Mellon University, 1995.
- [464] SMITH, M. und T. KANADE: *Video Skimming and Characterization through the Combination of Image and Language Understanding*. In: *IEEE International Workshop on Content-Based Access of Image and Video Databases*, S. 61–70. IEEE Computer Society Press, Januar 1998.
- [465] SMITH, M. A.: *Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 775–781. IEEE Computer Society Press, 1997.
- [466] SOBOTTKA, K. und I. PITAS: *Face Localization and Facial Feature Extraction Based on Shape and Color Information*. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Bd. 3, S. 483–486. IEEE Computer Society Press, September 1996.
- [467] SOILLE, P.: *Morphologische Bildverarbeitung*. Springer Verlag, Berlin, Heidelberg, New York, 1998.

- [468] SOMERS, G. und R. N. WHYTE: *Hand posture matching for Irish Sign language interpretation*. In: *Proceedings of the 1st international symposium on Information and communication technologies*, Bd. 49, S. 439–444. Trinity College Dublin, 2003.
- [469] SONG, B. und J. RA: *Automatic Shot Change Detection Algorithm Using Multi-stage Clustering for MPEG-Compressed Videos*. In: *Journal of Visual Communication and Image Representation*, Bd. 12(3), S. 364–385, September 2001.
- [470] SONKA, M., V. HLAVÁČ und R. BOYLE: *Image processing, analysis and machine vision*. Chapman and Hall, London, UK, 1993.
- [471] SONKA, M., V. HLAVÁČ und R. BOYLE: *Image processing, analysis and machine vision*. Thomson Learning Vocational, Florence, 2. Aufl., 1998.
- [472] SOULIE, F., F. VIENNET und B. LAMY: *Multi-modular neural network architectures: applications in optical character and human face recognition*. In: *International Journal of Pattern Recognition and Artificial Intelligence*, Bd. 7(4), S. 721–755, 1993.
- [473] SPIES, H. und H. SCHARR: *Accurate Optical Flow in Noisy Image Sequences*. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, Bd. I, S. 587–592. IEEE Computer Society Press, 2001.
- [474] SRINIVASAN, S., D. PETKOVIC und D. PONCELEON: *Towards Robust Features for Classifying Audio in the CueVideo System*. In: *Proceedings of the ACM international conference on Multimedia (Part 1)*, S. 393–400. ACM Press, 1999.
- [475] STEIGER, O., T. EBRAHIMI und D. SANJUAN: *MPEG-Based Personalized Content Delivery*. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Bd. 3, S. 45–48. IEEE Computer Society Press, September 2003.
- [476] STERNBERG, S. R.: *Grayscale morphology*. In: *Computer Vision, Graphics, and Image Processing*, Bd. 35 (3), S. 333–355, September 1986.
- [477] STRÖM, J., T. JEBARA, S. BASU und A. PENTLAND: *Real Time Tracking and Modeling of Faces: An EKF-Based Analysis by Synthesis Approach*. In: *Proceedings of the IEEE International Workshop on Modelling People*, S. 55–61. IEEE Computer Society Press, 1999.
- [478] STURMAN, D. J., D. ZELTZER und S. PIEPER: *Hands-on interaction with virtual environments*. In: *Proceedings of the 2nd annual ACM SIGGRAPH symposium on User interface software and technology*, S. 19–24. ACM Press, 1989.
- [479] SULL, S., J. KIM, Y. KIM, H. CHANG und S. LEE: *Scalable Hierarchical Video Summary and Search*. In: *Proceedings of IS&T/SPIE conference on Storage and Retrieval for Media Databases*, Bd. 3215, S. 553–561, 2001.
- [480] SUN, J., Y. HOTTA, Y. KATSUYAMA und S. NAOI: *Low resolution character recognition by dual eigenspace and synthetic degraded patterns*. In: *Proceedings of ACM workshop on Hardcopy document (HDP)*, S. 15–22. ACM Press, 2004.

- [481] SUN, J., Z. WANG, H. YU, F. NISHINO, Y. KATSUYAMA und S. NAOI: *Effective text extraction and recognition for WWW images*. In: *Proceedings of ACM symposium on Document engineering*, S. 115–117. ACM Press, 2003.
- [482] SUN, X., A. DIVAKARAN und B. S. MANJUNATH: *A Motion Activity Descriptor and Its Extraction in Compressed Domain*. In: *Lecture Notes In Computer Science*, Bd. 2195, S. 450–457. Springer-Verlag, 2001.
- [483] SUNDARAM, H. und S. CHANG: *Determining Computable Scenes in Films and their Structures Using Audio-Visual Memory Models*. In: *Proceedings of the 8th ACM international conference on Multimedia*, S. 95–104. ACM Press, 2000.
- [484] SUNDARAM, H. und S.-F. CHANG: *Condensing Computable Scenes using Visual Complexity and Film Syntax Analysis*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*. IEEE Computer Society Press, August 2001.
- [485] SUNDARAM, H. und S.-F. CHANG: *Constrained Utility Maximization for generating Visual Skims*. In: *Proceedings of 5th IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL)*, S. 124–131. IEEE Computer Society Press, Dezember 2001.
- [486] SUNDARAM, H. und S.-F. CHANG: *Video Analysis and Summarization at Structural and Semantic Levels*. In: FENG, D., W. C. SIU und H.-J. ZHANG (Hrsg.): *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*. Springer Verlag, März 2003.
- [487] SUNDARAM, H., L. XIE und S.-F. CHANG: *A Utility Framework for the Automatic Generation of Audio-Visual Skims*. In: *Proceedings of SIG ACM Conference On Multimedia*, S. 189–198. ACM Press, Dezember 2002.
- [488] SUNG, K.-K. und T. POGGIO: *Example-based learning for view-based human face detection*. Techn. Ber. A.I. Memo No. 1521, MIT, Cambridge, MA, USA, 1994.
- [489] SUZUKI, M., F. TAMARI, R. FUKUDA, S. UCHIDA und T. KANAHORI: *INFTY—An integrated OCR system for mathematical documents*. In: *Proceedings of ACM Symposium on Document Engineering*, S. 95–104. ACM Press, 2003.
- [490] SYEDA-MAHMOOD, T. und D. PONCELEON: *Learning video browsing behavior and its application in the generation of video previews*. In: *Proceedings of the 9th ACM international conference on Multimedia*, S. 119–128. ACM Press, 2001.
- [491] SZELISKI, R.: *Video mosaics for virtual environments*. In: *IEEE Computer Graphics and Applications*, Bd. 16(2), S. 22–30. IEEE Computer Society Press, März 1996.
- [492] SZELISKI, R. und H. SHUM: *Creating full view panoramic image mosaics and environment maps*. In: *Proceedings of Computer graphics and interactive techniques*, S. 251–258. ACM Press, 1997.
- [493] SÁNCHEZ, J., X. BINEFA, P. RADEVA und J. VITRIÀ: *Local Color Analysis for Scene Break Detection Applied to TV Commercials Recognition*. In: *Proceedings of International Conference on Visual Information and Information Systems (VISUAL)*, S. 237–244. Springer Verlag, Juni 1999.

- [494] TANIGUCHI, Y., A. AKUTSU und Y. TONOMURA: *PanoramaExcerpts: extracting and packing panoramas for video browsing*. In: *Proceedings of the 5th ACM international conference on Multimedia*, S. 427–436. ACM Press, 1997.
- [495] TARR, M. D. und H. H. BÜLTHOFF (Hrsg.): *Object Recognition in Man, Monkey, and Machine*. MIT Press, Cambridge, MA, USA, 1998.
- [496] TARR, M. J.: *Pattern recognition*. In: KAZDIN, A. (Hrsg.): *Encyclopedia of Psychology*. American Psychological Association, Washington, DC, USA, 2000.
- [497] TARR, M. J.: *Object Recognition*. In: NADEL, L. und R. GOLDSTONE (Hrsg.): *Encyclopedia of Cognitive Science*, S. 490–494. Nature Publishing Group/Macmillan Publishers Limited, London, UK, 2002.
- [498] TARR, M. J. und Q. C. VUONG: *Visual Object Recognition*. In: PASHLER, H. und S. YANTIS (Hrsg.): *Stevens' Handbook of Experimental Psychology: Sensation and Perception, Vol. 1*, S. 287–314. John Wiley and Sons, Inc., New York, NY, USA, 3. Aufl., 2002.
- [499] TEFAS, A., C. KOTROPOULOS und I. PITAS: *Variants of dynamic link architecture based on mathematical morphology for frontal face authentication*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 814–819. IEEE Computer Society Press, 1998.
- [500] TJONDRONEGORO, D., Y.-P. P. CHEN und B. PHAM: *Sports video summarization using highlights and play-breaks*. In: *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, S. 201–208. ACM Press, 2003.
- [501] TORR, P. und A. ZISSERMAN: *Feature Based Methods for Structure and Motion Estimation*. In: TRIGGS, B., A. ZISSERMAN und R. SZELISKI (Hrsg.): *Vision Algorithms: Theory and Practice*, Bd. 1883 d. Reihe *Lecture Notes in Computer Science*, S. 278–294, Berlin, Heidelberg, 1999. Springer Verlag.
- [502] TORRES, L. und E. J. DELP: *New trends in image and video compression*. In: *X European Signal Processing Conference*, September 2000.
- [503] TRAKA, M. und G. TZIRITAS: *Panoramic view construction*. In: *Signal Processing: Image Communication*, Bd. 18(6), S. 465–481, Juli 2003.
- [504] TRAZEGNIES, C., C. URDIALES, A. BANDERA und F. SANDOVAL: *Planar shape indexing and retrieval based on Hidden Markov Models*. In: *Pattern Recognition Letters*, Bd. 23 (10), S. 1143–1151, 2002.
- [505] TRIER, Ø., A. JAIN und T. TAXT: *Feature extraction methods for character recognition – a survey*. In: *Pattern Recognition*, Bd. 29 (4), S. 641–662, 1996.
- [506] TSALAKANIDOU, F., S. MALASSIOTIS und M. G. STRINTZIS: *A 2D+3D Face Authentication System Robust Under Pose and Illumination Variations*. In: *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA)*, September 2005.

- [507] TSENG, B. und C. LIN: *Personalized Video Summary using Visual Semantic Annotations and Automatic Speech Transcriptions*. In: *IEEE Workshop on Multimedia Signal Processing*, S. 5–8. IEEE Computer Society Press, Dezember 2002.
- [508] TSENG, B., C.-Y. LIN und J. R. SMITH: *Using MPEG-7 and MPEG-21 for personalizing video*. In: *IEEE Multimedia*, Bd. 11(1), S. 42–52. IEEE Computer Society Press, Januar 2004.
- [509] TSENG, B. und J. SMITH: *Hierarchical Video Summarization Based on Context Clustering*. In: *Proceedings of IS&T/SPIE conference on Internet Multimedia Management Systems IV*, S. 14–25, November 2003.
- [510] TSENG, B. L. und C.-Y. LIN: *Personalized Video Summary using Visual Semantic Annotations and Automatic Speech Transcriptions*. In: *IEEE International Workshop on Multimedia Signal Processing*, S. 5–8. IEEE Computer Society Press, Dezember 2002.
- [511] TSENG, B. L., C.-Y. LIN und J. R. SMITH: *Video Summarization and Personalization for Pervasive Mobile Devices*. In: *Proceedings of IS&T/SPIE conference on Storage and Retrieval for Media Databases*, Bd. 4676, S. 359–370, Januar 2002.
- [512] TURK, M.: *Gesture recognition*. In: JACKO, J. A. (Hrsg.): *Handbook of virtual environments: Design, Implementation, and Applications*, Kap. 9. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, USA, 2002.
- [513] TURK, M. und A. PENTLAND: *Face processing: Models for recognition*. In: *Proceedings of IS&T/SPIE conference on Intelligent Robots and Computational Vision VII: Algorithms and Techniques*, Bd. 1192, S. 22–32, November 1989.
- [514] TURK, M. und A. PENTLAND: *Eigenfaces for Recognition*. In: *Journal of Cognitive Neuroscience*, Bd. 3(1), S. 71–86, 1991.
- [515] TURK, M. und A. PENTLAND: *Face Recognition using Eigenfaces*. In: *IEEE Conference on Computing Vision and Pattern Recognition*. IEEE Computer Society Press, Juni 1991.
- [516] TUSCH, R., H. KOSCH und L. BÖSZÖRMÉNYI: *VIDEX: an integrated generic video indexing approach*. In: *Proceedings of ACM international conference on Multimedia*, S. 448–451. ACM Press, 2000.
- [517] UCHIHASHI, S. und J. FOOT: *Summarizing Video using a Shot Importance Measure and a Frame-Packing Algorithm*. In: *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Bd. 6, S. 3041–3044, 1999.
- [518] UCHIHASHI, S., J. FOOTE, A. GIRGENSOHN und J. BORECZKY: *Video Manga: Generating Semantically Meaningful Video Summaries*. In: *Proceedings of ACM Multimedia*, S. 383–392. ACM Press, 1999.
- [519] UEDA, H., T. MIYATAKE und S. YOSHIZAWA: *IMPACT: An Interactive Natural-motion-picture Dedicated Multimedia Authoring System*. In: *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology*, S. 343–350. ACM Press, April-Mai 1991.

- [520] ULLMAN, S.: *High-level Vision: Object Recognition and Visual Cognition*. MIT Press, Cambridge, MA, USA, 1996.
- [521] VALENTIN, D., H. ABDI, B. EDELMAN und A. J. O'TOOLE: *Principal Component and Neural Network Analyses of Face Images: What Can Be Generalized in Gender Classification?*. In: *Journal of Mathematical Psychology*, Bd. 41, S. 398–412, 1997.
- [522] VALENTIN, D., H. ABDI und A. J. O'TOOLE: *Categorization and identification of human face images by neural networks: A review of linear auto-associator and principal component approaches*. In: *Journal of Biological Systems*, Bd. 2, S. 413–429, 1994.
- [523] VALENTIN, D., H. ABDI, A. J. O'TOOLE und G. W. COTTRELL: *Connectionist models of face processing: A survey*. In: *Pattern Recognition*, Bd. 27, S. 1208–1230, 1994.
- [524] VEMURI, S., P. DECAMP, W. BENDER und C. SCHMANDT: *Improving Speech Playback Using Time-compression and Speech Recognition*. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, S. 295–302, 2004.
- [525] VERMAAK, J., P. PÉREZ, M. GANGNET und A. BLAKE: *Rapid Summarisation and Browsing of Video Sequences*. In: *Proceedings of British Machine Vision Conference (BMVC)*, S. 1–10, September 2002.
- [526] VETRO, A.: *MPEG-21 Digital Item Adaptation: Enabling Universal Multimedia Access*. In: *IEEE Multimedia*, Bd. 11(1), S. 84–87. IEEE Computer Society Press, Januar 2004.
- [527] VETRO, A., C. CHRISOPOULOS und H. SUN: *Video Transcoding Architectures and Techniques. An Overview*. In: *IEEE Signal Processing Magazine*, Bd. 20(2), S. 18–29. IEEE Computer Society Press, März 2003.
- [528] VETRO, A., T. CHRISTOPOULOS und T. EBRAHIMI: *Special Issue on Universal Multimedia Access*. In: *IEEE Signal Processing Magazine*, Bd. 20(2), S. 69–79. IEEE Computer Society Press, März 2003.
- [529] VETRO, A., T. HAGA, K. SUMI und H. SUN: *Object-based Coding for Long-term Archive of Surveillance Video*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Bd. 2, S. 417–420. IEEE Computer Society Press, 2003.
- [530] VETRO, A. und H. SUN: *An Overview of MPEG-4 Object-Based Encoding Algorithms*. In: *International Conference on Information Technology: Coding and Computing (ITCC)*, S. 366–369, April 2001.
- [531] V.WU, R.MANMATHA und E.M.RISEMAN: *TextFinder: An Automatic System to Detect and Recognize Text In Images*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 21, S. 1224–1229. IEEE Computer Society Press, November 1999.
- [532] WACTLAR, H.: *New Directions in Video Information Extraction and Summarization*. In: *DELOS Workshop*, S. 1–10, Juni 1999.
- [533] WACTLAR, H. D.: *Informedia – Search and Summarization in the Video Medium*. In: *Proceedings of Imagina*, S. 1–10, Januar 2000.

- [534] WACTLAR, H. D., M. G. CHRISTEL, Y. GONG und A. G. HAUPTMANN: *Lessons Learned From Building A Terabyte Digital Video Library*. In: *IEEE Computer*, Bd. 32(2), S. 66–73. IEEE Computer Society Press, 1999.
- [535] WANG, L., M. LEW und G. XU: *Offense based temporal segmentation for event detection in soccer video*. In: *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, S. 259–266. ACM Press, 2004.
- [536] WATT, A. und F. POLICARPO: *The Computer Image*. Addison-Wesley, Harlow, Essex, England, 1998.
- [537] WEICKERT, J.: *Anisotropic Diffusion in Image Processing*. European Consortium for Mathematics in Industry. Teubner, Stuttgart, 1998.
- [538] WEICKERT, J. und C. SCHNÖRR: *Optic Flow Calculation with Nonlinear Smoothness Terms Extended into the Temporal Domain*. Techn. Ber. TR-99-4, Department for Mathematics and Computer Science, University of Mannheim, Germany, 1999.
- [539] WEICKERT, J. und C. SCHNÖRR: *Variational optic flow computation with a spatio-temporal smoothness constraint*. In: *Journal of Mathematical Imaging and Vision*, Bd. 14(3), S. 245–255, Mai 2001.
- [540] WEINER, A. und M. MCGUIRE: *Google Announcement Sets Stage for Video Marketplace in 2006*. Techn. Ber. G00137334, Gartner Research, Januar 2006.
- [541] WENG, J. und D. L. SWETS: *Face Recognition*. In: JAIN, A. K., R. BOLLE und S. PANKANTI (Hrsg.): *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Press, Hingham, MA, USA, 1999.
- [542] WERBOS, P. J.: *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. Doktorarbeit, Harvard University, Cambridge, MA, USA, 1974.
- [543] WERNICKE, A. und R. LIENHART: *On the Segmentation of Text in Videos*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Bd. 3, S. 1511–1514. IEEE Computer Society Press, Juli 2000.
- [544] WEXELBLAT, A.: *An approach to natural gesture in virtual environments*. In: *ACM Transactions on Computer-Human Interaction (TOCHI)*, Bd. 2(3), S. 179–200. ACM Press, September 1995.
- [545] WILDEMUTH, B. M., G. MARCHIONINI, M. YANG, G. GEISLER, T. WILKENS, A. HUGHES und R. GRUSS: *How fast is too fast? Evaluating fast forward surrogates for digital video*. In: *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, S. 221–230. IEEE Computer Society Press, 2003.
- [546] WILDER, J., P. J. PHILLIPS, C. JIANG und S. WIENER: *Comparison of Visible and Infra-Red Imagery for Face Recognition*. In: *Proceedings of International Conference on Automatic Face and Gesture Recognition (ICAFGR)*, S. 182–187, 1996.

- [547] WINSCHER, L. und S. KOPF: *Entwicklung einer Börsensimulation mit der multiagentenbasierten Entwicklungsumgebung NetLogo*. Techn. Ber. TR-04-007, Department for Mathematics and Computer Science, University of Mannheim, Oktober 2004.
- [548] WISKOTT, L., J.-M. FELLOUS, N. KRÜGER und C. VON DER MALSBURG: *Face recognition by elastic bunch graph matching*. In: *Proceedings of International Conference on Computer of Images and Patterns (CAIP)*, Bd. 1296, S. 456–463, 1997.
- [549] WOLBERG, G.: *Digital Image Warping*. IEEE Computer Society Press, Los Alamitos, CA, 1990.
- [550] WREN, C. R., A. AZARBAYEJANI, T. DARRELL und A. PENTLAND: *Pfinder: Real-Time Tracking of the Human Body*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 19 (7), S. 780–785. IEEE Computer Society, Juli 1997.
- [551] WU, W., X. CHEN und J. YANG: *Incremental detection of text on road signs from video with application to a driving assistant system*. In: *Proceedings of ACM international conference on Multimedia*, S. 852–859. ACM Press, 2004.
- [552] XI, J., X.-S. HUA, X.-R. CHEN, L. WENYIN und H.-J. ZHANG: *A Video Text Detection and Recognition System*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, S. 873–876. IEEE Computer Society Press, 2001.
- [553] XIONG, Y. und K. TURKOWSKI: *Creating image-based VR using a self-calibrating fisheye lens*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 237–243. IEEE Computer Society Press, Juni 1997.
- [554] XU, C., X. SHAO, N. C. MADDAGE und M. S. KANKANHALLI: *Automatic music video summarization based on audio-visual-text analysis and alignment*. In: *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, S. 361–368. ACM Press, 2005.
- [555] XU, C., Y. ZHU und Q. TIAN: *Automatic music summarization based on temporal, spectral and cepstral features*. In: *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Bd. 1, S. 117–120. IEEE Computer Society Press, 2002.
- [556] YAGLOM, I. M.: *Geometric Transformations I (Number 8)*. Random House, New York, 1962.
- [557] YAHIAOUI, I., B. MÉRIALDO und B. HUET: *Automatic Video Summarization*. In: *Multimedia Content Based Indexing and Retrieval (MMCBIR)*, S. 1–4, September 2001.
- [558] YAHIAOUI, I., B. MÉRIALDO und B. HUET: *Optimal video summaries for simulated evaluation*. In: *Proceedings of European Workshop on Content-Based Multimedia Indexing (CBMI)*, S. 1–8, September 2001.
- [559] YAHIAOUI, I., B. MÉRIALDO und B. HUET: *Comparison of Multiepisode Video Summarisation Algorithms*. In: *Journal on Applied Signal Processing*, Bd. 1, S. 48–55. Hindawi Publishing Corporation, 2003.

- [560] YAN, H., Y. ZHANG, Z. HOU und M. TAN: *Automatic Text Detection In Video Frames Based on Bootstrap Artificial Neural Network And CED*. In: *International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision (WSCG)*, S. 1–6, Februar 2003.
- [561] YAN, W.-Q. und M. S. KANKANHALLI: *Detection and removal of lighting and shaking artifacts in home videos*. In: *Proceedings of the 10th ACM international conference on Multimedia*, S. 107–116. ACM Press, 2002.
- [562] YANG, G. und T. HUANG: *Human Face Detection in Complex Background*. In: *Pattern Recognition*, Bd. 27 (1), S. 53–63, 1994.
- [563] YANG, J., X. CHEN, J. ZHANG, Y. ZHANG und A. WAIBEL: *Automatic Detection and Translation of Text from Natural Scenes*. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Bd. 2, S. 2101–2104. IEEE Computer Society Press, Mai 2002.
- [564] YANG, J. und A. WAIBEL: *A real-time face tracker*. In: *Proceedings of IEEE Workshop on Applications of Computer Vision (WACV)*, S. 142–147. IEEE Computer Society Press, 1996.
- [565] YANG, M.-H. und N. AHUJA: *Detecting Human Faces in Color Images*. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Bd. 1, S. 127–130. IEEE Computer Society Press, 1998.
- [566] YANG, M.-H., D. J. KRIEGMAN und N. AHUJA: *Detecting faces in images: a survey*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 24 (1), S. 34–58. IEEE Computer Society Press, Januar 2002.
- [567] YANG, Y., K. SUMMERS und M. TURNER: *A text image enhancement system based on segmentation and classification methods*. In: *Proceedings of ACM workshop on Hardcopy document processing*, S. 33–40. ACM Press, 2004.
- [568] YE, M. und R. M. HARALICK: *Optical Flow From A Least-Trimmed Squares Based Adaptive Approach*. In: *Proceedings of International Conference on Pattern Recognition (ICPR)*, Bd. 3, S. 1052–1055, 2000.
- [569] YEO, B.-L. und B. LIU: *Rapid scene analysis on compressed video*. In: *IEEE Transactions on Circuits and Systems for Video Technology*, Bd. 5(6), S. 533–544. IEEE Computer Society Press, Dezember 1995.
- [570] YEO, B.-L. und M. YEUNG: *Retrieving and Visualizing Video*. In: *Communications of the ACM*, Bd. 40(12), S. 43–52. ACM Press, Dezember 1997.
- [571] YEUNG, M.: *Video Browsing Using Clustering and Scene Transitions on Compressed Sequences*. In: *Proceedings of IS&T/SPIE conference on Multimedia Computing and Networking*, Bd. 2417, S. 399–413, 1995.
- [572] YEUNG, M. M., B.-L. YEO und B. LIU: *Extracting story units from long programs for video browsing and navigation*. In: *Proceedings of IEEE International Conference on Multimedia Computing and Systems*, S. 296–305. IEEE Computer Society Press, 1996.

- [573] YOON, K. und S. B. JUN: *Real-time video indexing and non-linear video browsing for digital TV receivers with persistent storage*. In: *IEEE International Conference on Consumer Electronics (ICCE)*, S. 28–29. IEEE Computer Society Press, 2003.
- [574] YOW, K. C. und R. CIPOLLA: *Feature-based human face detection*. In: *Image Vision Computing*, Bd. 15(9), S. 713–735, 1997.
- [575] YU, B. und S. CAI: *A domain-independent system for sketch recognition*. In: *Proceedings of international conference on Computer graphics and interactive techniques*, S. 141–146. ACM Press, 2003.
- [576] YU, B., W.-Y. MA, K. NAHRSTEDT und H.-J. ZHANG: *Video summarization based on user log enhanced link analysis*. In: *Proceedings of the 11th ACM international conference on Multimedia*, S. 382–391. ACM Press, 2003.
- [577] YU, K., X. JIANG und H. BUNKE: *Face Recognition by Facial Profile Analysis*. In: *Proceedings of International Workshop on Automatic Face- and Gesture-Recognition (IWAAGR)*, S. 208–213, 1995.
- [578] YUILLE, A. L., P. W. HALLINAN und D. S. COHEN: *Feature extraction from faces using deformable templates*. In: *International Journal of Computer Vision*, Bd. 8(2), S. 99–111. Kluwer Academic Publishers, August 1992.
- [579] ZABIH, R., J. MILLER und K. MAI: *A feature-based algorithm for detecting and classifying scene breaks*. In: *Proceedings of ACM International Conference on Multimedia*, S. 189–200. ACM Press, 1995.
- [580] ZABIH, R., J. MILLER und K. MAI: *Feature-Based Algorithms for Detecting and Classifying Scene Breaks*. Techn. Ber., Computer Science Department, Cornell University, Juli 1995.
- [581] ZABIH, R., J. MILLER und K. MAI: *A feature-based algorithm for detecting and classifying production effects*. In: *Multimedia Systems*, Bd. 7 (2), S. 119–128. Springer Verlag, 1999.
- [582] ZAHN, C. T. und R. Z. ROSKIES: *Fourier descriptors for plane closed curves*. In: *IEEE Transactions on Computers*, Bd. C–21(3), S. 269–281. IEEE Computer Society Press, 1972.
- [583] ZHANG, D. und S.-F. CHANG: *General and Domain-specific Techniques for Detecting and Recognizing Superimposed Text in Video*. In: *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Bd. 1, S. 593–596. IEEE Computer Society Press, 2002.
- [584] ZHANG, H. J., A. KANKANHALLI und S. SMOLIAR: *Automatic Partitioning of Full-Motion Video*. In: *Multimedia Systems*, Bd. 1 (1), S. 10–28, 1993.
- [585] ZHANG, J., Y. YAN und M. LADES: *Face Recognition: Eigenface, Elastic Matching, and Neural Nets*. In: *Proceedings of the IEEE*, Bd. 85(9), S. 1423–1435. IEEE Computer Society Press, September 1997.
- [586] ZHANG, Z., R. DERICHE, O. FAUGERAS und Q.-T. LUONG: *A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry*. In: *Artificial Intelligence: Special volume on computer vision*, Bd. 78(1–2), S. 87–119. Elsevier Science Publishers Ltd., Oktober 1995.

- [587] ZHAO, W., R. CHELLAPPA, P. J. PHILLIPS und A. ROSENFELD: *Face recognition: A literature survey*. In: *ACM Computing Surveys (CSUR)*, Bd. 35(4), S. 399–458. ACM Press, Dezember 2003.
- [588] ZHONG, H., J. SHI und M. VISONTAI: *Detecting unusual activity in video*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Bd. 2, S. 819–826. IEEE Computer Society Press, Juni 2004.
- [589] ZHONG, Y., H. ZHANG und A. K. JAIN: *Automatic Caption Localization in Compressed Video*. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Bd. 22 (4), S. 385–392. IEEE Computer Society Press, April 2000.
- [590] ZHU, X., J. FAN, A. K. ELMAGARMID und X. WU: *Hierarchical video content description and summarization using unified semantic and visual similarity*. In: *Multimedia Systems*, Bd. 9(1), S. 31–53, 2003.
- [591] ZHU, X., X. WU, J. FAN, A. K. ELMAGARMID und W. G. AREF: *Exploring video content structure for hierarchical summarization*. In: *Multimedia Systems*, Bd. 10(2), S. 98–115, 2004.
- [592] ZOGHLAMI, I., O. FAUGERAS und R. DERICHE: *Using geometric corners to build a 2D mosaic from a set of images*. In: *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, S. 420–425. IEEE Computer Society Press, Juni 1997.
- [593] ZUO, F. und P. H. DE WITH: *Automatic Human Face Detection for a Distributed Video Security System*. In: *PROGRESS Workshop 2002*, S. 269–274, Oktober 2002.
- [594] ZUO, F. und P. H. N. DE WITH: *Fast human face detection using successive face detectors with incremental detection capability*. In: *Proceedings of IS&T/SPIE conference on Image and Video Communications and Processing*, Bd. 5022, S. 831–841, Januar 2003.
- [595] ZUO, F. und P. H. N. DE WITH: *Fast Facial Feature Extraction Using a Deformable Shape Model with Haar-Wavelet Based Local Texture Attributes*. In: *International Conference on Image Processing (ICIP)*, S. 1425–1428, Oktober 2004.
- [596] ZUO, F. und P. H. N. DE WITH: *Multistage Facial Feature Extraction for Accurate Face Alignment*. In: *Proceedings of IS&T/SPIE conference on Visual Communications and Image Processing (VCIP)*, Bd. 5308, S. 773–781, Januar 2004.
- [597] ZUO, F. und P. H. N. DE WITH: *Real-Time Facial Feature Extraction by Cascaded Parameter Prediction and Image Optimization*. In: *International Conference on Image Analysis and Recognition*, Bd. 3212, S. 651–659, Oktober 2004.

Index

A

Abstand zur Kamera	214
Adaption	
Audiosignal	149
Bild	149
Bildauflösung	146, 150, 156
Bildqualität	164
Bildwiederholrate	150
Bitrate	150
Client	147
Echtzeit	147
Farbtiefe	146, 150, 155
Multimediale Inhalte	147
Proxy	147
Server	147
Statisch	147
Video	145 f, 149
Zeitpunkt	147
Additive dissolve	12
Änderungen einer Kontur	210
Affine Transformation	40
Aggregation von Ergebnissen	210
Aggregierter Merkmalswert	188
Amateurvideo	146, 167, 179
Amplitudenmodulierte Raster	153
Anpassung	
Abspielgeschwindigkeit	149
Helligkeit	164
Kontrast	165
Bildgröße	149
Attention object	149
Audio	196
Audioadaption	149
Audioanalyse	188
Ausblendung	9
Ausschneiden von Bildbereichen	156

Auswahl einer Bildregion	156
Authentifizierung eines Gesichtes	127

B

Background-Sprites	63
Base layer	148
Basisschicht	148
Basisvektor	138
Belichtung	150
Bewegungsaktivität	177, 187, 191
Bewegungsanalyse	205
Fahrzeug	212
Person	215
Bewegungsvektor	41
Bildadaption	149
Bildauflösung	145 f
Bildfehler	34, 150
Bildqualität	33, 164
Bildtransformation	57
Bildwiederholrate	146, 181
Bitrate	146
Blockmatching-Verfahren	42
Brennweite	215
Buchstabentrenner	110

C

Camera motion	39
Canny-Kantendetektor	19
Canonical view	72
Chrominanz	59
Closing-Operator	64
Compactness	75
Compression network	133
Connectionist model	130
Content repurposing	147
Cropping	156

Cross dissolve 12
 CueVideo-System 181
 Curvature scale space image 77
 Cut 7
 Cylindrical camera model 40

D

DCT-Koeffizient 107
 Dialog 8, 186
 Digital item adaptation 148
 Dijkstra-Algorithmus 112
 Dilatation 20, 64
 Dissolve 9
 Dolly shot 40
 Dominante Farbe 214
 Drucktechnik 153
 Dynamische Adaption 147
 Dynamische Programmierung 211

E

Earth-Movers-Distanz 35 f
 Eccentricity 75
 Ecke 41
 Edge change fraction 19
 Edge change ratio 19
 Edge-based contrast 21
 Eigenbild 131
 Eigenface 131
 Eigengesicht 131, 138
 Eigenpicture 131
 Eigenvektor 131, 138
 Einblendung 9
 Einzelbild 8
 Enhancement layer 148
 Erkennung
 Gesten 205
 Körperhaltung 215
 Erosion 64
 Error diffusion 153
 Euklidische Norm 15
 European Chronicles Online 33
 Projekt 100, 164, 174, 181, 201
 Evolution einer Kontur 76
 Exzentrizität 75

F

F1-Maß 24

Face

Detection 127
 Recognition 127
 Space 131
 Fade in 9
 Fade out 9
 Farbe 214
 Kleidung 216
 Farbraum 59
 Farbtiefe 145 f
 Reduktion 149
 Farbton 59
 Filmarchiv 33
 Frame 8
 Frequenzmodulierte Raster 153

G

Gaußglättung 17
 Genre eines Videos 178
 Geometrisch invariante Faktoren 76
 Geometrische Konturdeskriptoren 75
 Geräteklasse 145
 Gerätemerkmale 148
 Geschwindigkeit eines Objektes 215
 Gesicht 187, 191
 Normalisierung 137
 Gesichtsausdruck 127
 Gesichtserkennung 125, 127, 138
 Deformierbare Templates 129
 Dreidimensionales Modell 130
 Farbanalyse 129
 Gesichtsmerkmale 129
 Globale Merkmale 129
 Kantenanalyse 129
 Konnektionistische Verfahren ... 128, 130
 Modellbasierte Verfahren 128
 Profilanalyse 130
 Statische Templates 129
 Texturanalyse 129
 Gesichtsmerkmal 127 f
 Gesichtsraum 131, 138
 Gesichtsregion 127, 134, 157
 Gespiegelte Kontur 81
 Gesten 205
 Glättung einer Kontur 76
 Globale Konturdeskriptoren 75

Gradient descent 46
 Gradientenabstiegsverfahren 46
 Graphic text 106
 Greedy-Algorithmus 42

H

Halbton 153
 Hard cut 8
 Harris-Eckendetektor 41
 Harter Schnitt 8
 Hauptkomponentenanalyse 131
 Helligkeit 59
 Korrektur 164
 Lineare Transformation 151
 Schwankung 34, 164
 Helligkeitsschwankung 164
 Hintergrundbild 59, 180
 Histogramm 14
 Differenz 15
 Kumuliert 35, 151
 Historischer Film 33, 146
 Historisches Video 33, 100
 Hitchcock-System 181
 Hotelling-Transformation 131
 HSI-Farbraum 59
 Hue 59
 Hysterese 19

I

Impulsrauschen 120
 Informedia-System 181
 Intelligente Räume 206
 Intensity 59

K

K-Means-Algorithmus 62, 115, 180
 Kamera
 Bewegung 39
 Operation 39
 Parameter 42
 Rotation 40
 Kamerabewegung 186, 189 f
 Dauer 190
 Stärke 190
 Kameraeinstellung 7
 Ähnliche Gruppen 184

Ähnlichkeit 184, 192
 Auswahl 178, 188, 194
 Repräsentative Bilder 173, 179, 183
 Überbelichtet 164
 Unterbelichtet 164
 Verwackelt 150, 164, 166
 Kamerafahrt 40
 Kameramodell 40
 Sphärisch 40
 Zylindrisch 40
 Kameraüberwachung 206
 Kanonische Sicht 72, 90
 Kantenänderungsrate 19
 Ausgehende Kantenpixel 19
 Eingehende Kantenpixel 19
 Kantenbasierter Kontrast 19, 21
 Kantenbild 153
 Kantendetektor 19
 Karhunen-Loève-Transformation 131
 Key frame 150, 173
 KidsRoom 206
 Kleinste getrimmte Quadrate 43
 Kollage 198
 Kompaktheit 75
 Komprimierendes Netzwerk 133
 Konnektionistisches Modell 130
 Kontrast 192
 Kontur 72
 Vergleich 79
 Konturprofil 117
 Konvexe Objektregion 84
 Konvexes Objekt 84
 Korrelationsmatrix 180
 Kovarianzmatrix 131
 Kratzer 34
 Korrektur 164 f
 Krümmung 77
 Krümmungsbasierter Skalenraum 76
 Künstliche Kamerabewegung 162
 Kürzeste-Pfade-Algorithmus 112
 Kürzester Pfad im Graph 211

L

L_1 -Norm 15, 35
 L_2 -Norm 15, 35
 Least trimmed squares 43

Linear autoassociative network 131
 Lineare Interpolation 58
 Lineares autoassoziatives Netz 131
 Linie 150
 Linienpixel 166

M

Maß für die Zuverlässigkeit 88
 Median 59, 63
 Metadaten 174
 Minimal perceptible size 159
 Minkowski-Metrik 15
 Mobiles Gerät 145
 MoCA-Projekt 178, 181
 Morphing 130
 Morphologischer Operator 64
 Motion vector 41
 Motion-Activity-Deskriptor 177, 187
 MPEG-7 148
 MPEG-21 148
 Musikvideo 179

N

Nachrichtensendung 178
 Navigation innerhalb einer Videos 173
 Neural net 131
 Neuronales Netz 131
 Ausgabeschicht 132
 Eingabeschicht 132
 Verdeckte Schicht 132
 Nichtlineares autoassoziatives Netz 131 f
 Nullstellen der Krümmungsfunktion 76 f
 Nutzer
 Anfrage 148
 Historie 148
 Präferenz 148

O

Object motion 39
 Objekt 187, 191
 Beschreibung 209
 Bewegung 39
 Bezeichnung 71
 Farbe 214
 Position 214
 Region 158

Segmentierung 63
 Objektänderung 205
 Objekterkennung 71
 Differenz zu Objekten 89
 Historische Videos 100
 Mehrdeutigkeit 83
 Vergleich verrauschter Objekte 82
 Objektklasse 71
 Differenz zum Objekt 211
 Wechsel 211
 OCR 105
 Offsetdruck 153
 Opening-Operator 64
 Optical flow 41
 Optical character recognition 105
 Optimaler Pfad im Graph 212
 Optischer Fluss 41
 Orts-Zeit-Bild 14

P

Pan 40
 Panoramabild 59, 63, 180
 Parametrisierung einer Kontur 74
 Pattern-Matching 116, 136
 Personalisierung 148
 Präzision 23
 Precision 23
 Principal component analysis 131
 Projection profile 109
 Projektionsprofil 109 f, 136

R

Rasterung 153
 Rauschen 33
 Recall 23
 Referenzbild 167
 Region innerhalb einer Kontur 87
 Region of interest 149
 Region-Growing-Algorithmus 114
 Region-Merging-Algorithmus 113
 Reißschwenk 49
 Robuste Regressionsschätzung 43
 Rotationsinvarianter Konturvergleich 81

S

Sättigung 59

- Salient point 41
 Salt and pepper noise 120
 Saturation 59
 Scaling 156
 Scene 8
 Scene text 106
 Schnitt 7 f
 Schnitterkennung 7, 11
 Schwarz-Weiß-Filme 33
 Schwellwert
 Absolut 11
 Adaptiv 11
 Schwenk 40
 Schwerpunkt
 Konturpixel 75
 Objekt 214
 Segmentierung 55, 63
 Buchstabe 110
 Gesicht 135
 Objekt 55
 Semantik 1
 Semantische Transkodierung 148
 Semantisches Merkmal 157
 Bewertung 158
 Informationsgehalt 159
 Serie 179
 Shape-Contexts 117
 Shot 7
 Singulärwertzerlegung 180
 Singular value decomposition 180
 Skalenraumabbildung 77, 117
 Ähnlichkeit 81
 Bogen 78
 Breite eines Bogens 83
 Differenz der Bögen 80
 Konvexe Regionen 79
 Merkmale 82
 Signifikante Bögen 94
 Zuordnung von Bögen 80
 Skalierung 156
 Video 148
 Smart room 206
 Soft cut 8
 Sparse features 41
 Spherical camera model 40
 Spielfilm 178
 Sportvideo 178
 Spracherkennung 149
 Standardabweichung 17
 Streifen
 im Bild 34
 Korrektur 164 f
 Struktur eines Videos 176
 Strukturelement 64
 Summe absoluter Differenzen 13
 Superimposed text 106
 Support-Vector-Maschine 132
 Surveillance 206
 Swish pan 49
 Szene 8, 184 f, 192
 Szenentext 106

T
 Text
 Farbe 113
 Pixel 113
 Region 109
 Texterkennung 105, 116
 in Bildern und Videos 120
 Regelbasiert 107
 Regionenbasiert 107
 Texturbasiert 107
 Textregion 158
 Textur 155
 Tilt 40
 Tonhöhe 181
 Trailer 175
 Transformation 48, 56 f
 Farbbild 59
 Helligkeit 151
 Kontur 84
 Transkodierung eines Videos 148 f
 TRECVID-Konferenz 22

U
 Überblendung
 Additiv 12
 Kreuz 12
 Übergangsmatrix 211
 Überlagerter Text 106
 Überblendung 9
 Überwachungssystem 206
 Universal multimedia access 148

Unterklasse	209
Urlaubsvideo	179
Usage environment description	148
User preference description	148

V

VAbstract	178
Verbesserung der Bildqualität	146, 164
Verwackeltes Video	34
Video	
Abstract	174
Skim	174
Summary	174
Surveillance	127
Video-Zusammenfassung	173 – 176, 181
Dynamisch	180, 201
Statisch	179, 198
Videoüberwachung	127
Videoadaption	145 f
Videoarchiv	173
Videokollage	198
Videonavigation	173
Visual descriptor	72
Visuelle Komplexität	177
Visueller Deskriptor	72
Vollständigkeit	23
Vorschau eines Videos	175

W

Wahrnehmungsebene eines Objektes	71
Weicher Schnitt	8
Wipe	9
Wischeffekt	9

Y

YUV-Farbraum	59
--------------------	----

Z

Zeichenerkennung	105
Zeichensprache	206
Zoning	116
Zoom	40
Zoom-in	40
Zoom-out	40
Zusammenfassung eines Videos	173